RICE UNIVERSITY

Assessing Adverse Impact:
An Alternative to the Four-Fifths Rule

by

Seydahmet Ercan

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Arts

APPROVED, THESIS COMMITTEE

Frederick L. Oswald, Professor, Chair,
Psychology

Margaret E. Beier, Associate Professor,
Psychology

Anton J. Villado, Assistant Professor,
Psychology

HOUSTON, TEXAS
April 2012

ABSTRACT

Assessing Adverse Impact:

An Alternative to the Four-Fifths Rule

by

Seydahmet Ercan

The current study examines the behaviors of four adverse impact measurements: the 4/5ths rule, two tests of significance (Z_D and Z_{IR}), and a newly developed AI measurement (Ln_{adj}). Upon the suggestion of the Office of Federal Contract Compliance Program Manual about the sensitivity of the assessment of AI when the sample size is very large (Office of Federal Contract Compliance Programs, 2002), Ln_{adj} is a new statistic that has been developed and proposed as an alternative practical significance test to the 4/5ths rule. The results indicated that, unlike the 4/5ths rule and other tests for adverse impact, Ln_{adj} is an index of practical significance that is less sensitive to differences across selection conditions that are not supposed to affect tests of adverse impact. Furthermore, Ln_{adj} decreases Type I error rates when there is a small *d* value and Type II error rates when there is moderate to large *d* value.

Acknowledgements

First, I would like to thank my parents and other family members for their words of encouragement they have given during my graduate school. I would also like to thank my academic advisor Fred Oswald for his invaluable insight, guidance, and encouragement. I am also grateful to my committee members, Margaret Beier and Anton Villado, for their insightful comments and feedbacks.

Lastly, I wish to express sincere appreciation to Turkish Ministry of National Education for their financial support during my graduate school.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

When making a selection or promotion decision, one of the most important challenges a selection decision maker has to address is identifying whether or not the selection procedure at hand has an adverse impact (AI) on minority groups. A great deal of previous research has discussed the appropriateness of different selection strategies depending on the relative value firms place on performance or on minority representation (Cascio, Outtz, Zedeck, & Goldstein, 1995; De Corte, Lievens, & Sackett, 2007; Hattrup, Rock, & Scalia, 1997; Hunter, Schmidt, & Rauschenberger, 1977; Pulakos & Schmitt, 1996; Sackett & Ellingston, 1997; Sackett & Roth, 1996). Although there is no question that the general goal of putting effort, time, and money to develop and implement a selection procedure (instead of relying on random selection) is to hire personnel who have the required KSAOs (knowledge, skills, abilities, and other characteristics) to perform well on the job, this research has also placed a significant priority on minority representation and a lack of AI as desired personnel selection outcomes.

Although little-to-no AI is a desirable selection outcome that promotes diversity in the workplace, the most valid selection tools are those that measure cognitive ability and, unfortunately, tend to result in higher level of AI. For example, Hunter and Hunter (1984), reanalyzing Ghiselli's (1973) work on mean validity of several selection tools and Hunter's (1981) meta-analyses of the United States Employment Service data base of 515 validation studies, found that general cognitive ability test was the most valid predictor of job performance for the all job families studied except vehicle operator job family for

which the general psychomotor ability test was found the most valid predictor. The results of the meta-analysis by Hough, Oswald, and Ployhart (2001), however, showed that general cognitive ability tests had large AI on racial minority groups except East Asians. Therefore, AI becomes an inevitable concern for the organizations that want to make use of the most valid selection tools. This situation has been addressed in the literature as the diversity-validity dilemma (Ployhart & Holtz, 2008; Pyburn, Ployhart, & Kravitz, 2008; Theron, 2009). It is a dilemma because organizations, regardless of business or legal reasons, want to attain both higher validity and lesser AI, yet there is a necessary tradeoff between the two when selection measures related to cognitive ability are involved.

Considering the negative legal consequences of using a selection procedure causing AI, it is not surprising that organizations want to attain higher minority representation or lesser AI. Being obliged to pay the monetary compensation awarded for the plaintiff is one of these legal consequences. Although the amount must be directly related to the actual monetary losses by the plaintiff, there is no limit on the expenses that companies may be required to cover when they lose the lawsuit. In addition, companies are required to pay monetary reward to each plaintiff for punitive damages and damages for emotional distress. The ceiling for these extra damages is between $50,000 and $300,000 for each reward, depending on the size of the company (Landy, 2005). Covering the sum of these actual and extra damages can be quite devastating for the financial standing of a company. Thus, replacing a selection procedure with an equally valid alternative that mitigates AI is particularly important to reduce (and hopefully remove) potentially expensive lawsuits to be filed by members of the protected groups

that are otherwise affected adversely from selection procedures. Even if there is no alternative selection procedure, being aware beforehand that AI is a possible outcome, organizations can take proactive measures in a timely manner to defend themselves in a possible court case. Some of these proactive measures include use of measures of additional relevant constructs (e.g., personality or interpersonal skills), use of coaching or test orientation programs when performance to a standard is an issue (as in the licensure or certification exams), and use of alternate modes of presenting test stimuli (Sackett, Schmitt, Ellingson, & Kabin, 2001).

In addition to the financial compensation that organizations with AI in their selection procedures may be legally forced to pay, being sued for using a test causing AI may also increase negative public perceptions and damage the prestige of the organization, because these cases are often debated extensively in the popular media (e.g., The *Ricci et al. v. DeStefano et al.* case was extensively discussed in *The New York Times*, *Forbes*, *The Washington Post* and *The Wall Street Journal*[1]). Even though it is difficult to put a price on prestige lost, the cost may be well over the amount that the organization can cover, especially if the market is highly competitive. This is because profitability decreases (Rumelt, 1991) and consumers can easily switch to a very similar (if not the same) product offered by other companies in a highly competitive marketplace. In other words, losing prestige easily leads to losing customers in competitive markets, which in turn might lead to a cost that those organizations with their low profitability fall short of being able to cover it. Thus, assessing selection procedures to identify those that

[1]http://www.nytimes.com/2009/06/30/us/30scotus.html
http://online.wsj.com/article/SB124631901145470941.html
http://www.washingtonpost.com/wp-dyn/content/article/2009/06/29/AR2009062901608.html
http://www.forbes.com/2009/06/30/ricci-destefano-supreme-court-opinions-contributors-connecticut-firefighters- race.html

cause AI and removing or being prepared to defend them in the court are important steps that should be followed by organizations as a precaution to prevent potentially expensive lawsuits and prestige lost.

Organizations are obviously not the only side affected by the consequences of using a selection procedure causing AI. In addition to the aforementioned hardships that organizations face, minority applicants might be more negatively affected by these selection procedures than majority-group applicants.

In the context of Equal Employment Opportunity Program (EEOP), *minority* refers to the group of people within a country or state that differs from the dominant group in terms of their race/color, religion and nationality (EEOP, n.d, para. 23). Because the ratio of females to males is around 1.00 in the USA[2], sex does not define minority status. However, Equal Employment Opportunity Commission (EEOC) requires organizations to maintain records of employment decision by sex and following races and ethnic groups: Blacks, American Indians (including Alaskan Natives), Asians (including Pacific Islanders), Hispanic (including persons of Mexican, Puerto Rican, Cuban, Central or South American, or other Spanish origin or culture regardless of race), Whites (Caucasians) other than Hispanic, and total (EEOC et al., 1978). Besides, EEOP makes it clear that women are considered having "minority status" because they have been systematically excluded from the economy as have various minorities (EEOP, n.d, para. 23). Thus, *minority* mainly refers to any of these groups (including women) throughout the paper[3].

---

[2] http://nationalatlas.gov/articles/people/a_gender.html

[3] Although EEOC requires to maintain employment decisions by sex and the above presented races and ethnic groups, discrimination types are not limited to sex, race/color, and ethnic background. The other types of discrimination prohibited by the laws and enforced by EEOC include age, disability, equal

These minority applicants as well as women might be unduly affected by the selection test with AI to the extent that there are equally or more valid selection measures with reduced or no AI. For example, previous research indicated that a content valid test of job sample developed to measure KSAOs for a specific technical area causes less AI impact than a content valid written achievement test developed to measure the same KSAOs (Schmidt, Greenthal, Hunter, Berner, & Seaton, 1977). When there is a test of job sample available as a valid alternative, using a written achievement test will adversely affect minority applicants and unduly reduce their chance of being hired. This is because a written achievement test requires reading comprehension ability that might be or not be job related yet increases the cognitive loading of the test (Chan & Schmitt, 1997; Sackett et al., 2001). Increasing the cognitive loading of a test unnecessarily will lead to adversely affecting minority applicants (e.g., Hispanics or Blacks) who usually perform poorer than majority applicants (White Americans) on the cognitive ability tests. Thus, assessing various selection procedures and using those with reduced or no AI has important implications to prevent minority applicants[4] from being unduly affected by the selection procedure.

---

pay/compensation, genetic information, national origin, pregnancy, religion, retaliation, and sexual harassment (EEOC, discrimination by type obtained from http://www.eeoc.gov/laws/types/index.cfm).

[4] In discussing minorities being unduly affected by a selection procedure with AI, it is important not to discuss as if all minorities are the same. There are 5 major minority groups addressed in the latest US census: Blacks (12.6%), American Indians and Alaska Natives (.9%), Asians (4.8%), Native Hawaiian and other Pacific Islanders (.2%), and Hispanics/Latinos (16.3%; U.S. Census Bureau, 2010). The recent report by U.S. Department of Labor and U.S. Bureau of Labor Statistics (2011) presented that each of the largest three minority groups constitutes more than 2% of the total labor force: Hispanics/Latinos, Blacks, and Asians constitute respectively about 13%, 10%, and 4% of the labor force. Knowing that *Uniform Guidelines* recommended including each group which constitutes at least 2% of the labor force in a relevant labor area into AI analysis, Hispanics/Latinos, Blacks, and Asians are more likely to be observed as minority groups in any selection scenario. Previous research clearly demonstrated that the magnitude of AI, as a result of a selection test, on each of these minority groups were not the same. For example, if the cognitive loading of the selection test is high, Blacks are affected most adversely, Hispanics/Latinos are affected less adversely than Blacks, and East Asians are not affected adversely at all. This is because Black-White and Hispanics/Latinos-White subgroup mean differences on cognitive ability are respectively 1.00

Underlying the importance of assessing whether or not a selection procedure has AI on minority groups, the next and fundamental question inevitably becomes how to asses AI. *Uniform Guidelines on Employee Selection Procedures* (referred as *Uniform Guidelines* from now on) recommended assessing AI by comparing the selection rate of the group with the lowest selection ratio to the selection rate of the group with the highest selection ratio (EEOC et al., 1978). If the selection rate of the group with the lowest selection ratio is less than 4/5ths of the selection rate of the group with the highest selection ratio, it indicates AI. Although *Uniform Guidelines* recommended using this practical procedure, known as the 4/5ths rule, to assess AI, the practicality of using this rule is not without question (Bobko, Roth, & Potosky, 1999; Greenberg, 1979; Morris & Lobsenz, 2000). A quantitative analysis by Greenberg (1979) indicated that application of the 4/5ths rule to detect AI produced large amount of Type I errors (i.e., concluding that there is AI on minority groups when both minority and majority groups have the same probability of passing) and Type II errors (i.e., concluding that no AI on minority group exists when in fact it is evident).

Even though the 4/5ths rule is described as the rule of thumb to assess AI, EEOC et al. (1978) advised organizations to use significance testing where large number of hiring is made. Statistical analysis had been considered as a helpful decision making tool by the Supreme Court in discrimination cases even before the *Uniform Guidelines* was introduced. The Supreme Court acknowledged the value of using tests of significance to determine if there is a significant difference between minority and majority group representation in *Castaneda v. Partida* (1977) and in *Hazelwood School District v. United*

---

and .50 (in favor of Whites) while Asian-White subgroup mean difference is .20 (in favor of Asians; Hough et al., 2001). Therefore, when discussing AI on minorities it is important to keep in mind that there are more than one minority group and each of them are affected differently from a selection procedure with AI.

States (1977) cases. Other than these, the Office of Federal Contract Compliance Program Manual (Office of Federal Contract Compliance Programs, 2002) recommended significance testing as a supplemental analysis to the 4/5ths rule when assessing AI. The OFCCP (Office of Federal Contract Compliance Programs) Manual even recommended using tests of practical and statistical significance rather than the 4/5ths rule when the sample size is very large. Whether or not the Manual refers to the 4/5ths rule by test of practical significance is, however, unclear. Yet there is no test of practical significance, other than the 4/5ths rule, which is suggested or recommended by the *Uniform Guidelines*, the OFCCP Compliance Manual, or the court.

Although the discussion in the previous paragraph indicates that statistical significance test has been considered as an informative decision making tool by the Supreme Court and the Federal Agencies, a recent decision reached by the Supreme Court in *Matrixx Initiatives, Inc., et al v. Siracusano et al*. (2011) case has a potential influence on the practice of using significance test as a decision-aid tool in employment decision cases. In this case, the investors who bought Matrixx stock between 2003 and 2004 alleged that Matrixx violated federal securities law by failing to disclose the information that there might be a possible link between use of Zicam, a cold remedy developed by Matrixx, and loss of the sense of smell. Although there were some customers complaining that they experienced loss of sense of smell after they used Zicam, Matrixx Initiative, Inc. argued that there was no need to disclose this information because none of the reports the company received showed statistically significant evidence confirming the claim of the consumers. After considering the argument by *Matrixx Initiatives, Inc.,* the Supreme Court concluded that companies cannot rely solely

on statistical significance in making their decision regarding which information they need to disclose to investors and stated that “something more than the mere existence of adverse event reports is needed to satisfy that standard, but that something more is not limited to statistical significance and can come from the source, content, and context of the reports” (*Matrixx Initiatives, Inc. et al v. Siracusano et al*., 2011, p.2). Therefore, relying solely on statistical significance by plaintiffs/defendants in AI assessment to support their position has the potential of being considered as an inconclusive approach by the Supreme Court.

Regardless of the recent decision reached by the Supreme Court, even if the legal and employment stakeholders consider significance testing as a viable alternative or supplement to the 4/5ths rule, they need to make a decision about which test of significance to use. Collins and Morris (2008) discussed four statistical difference tests: the *Z*-test on the difference between two proportions (the Z_D test), Fisher's Exact Test, Yates' continuity-corrected chi-square test, and a corrected chi-square test suggested by Upton (1982). Other than these tests, Morris and Lobsenz (2000) proposed a significance test (the Z_{IR} test) which is based on the selection ratio difference; and Biddle and Morris (2011) suggested using Lancaster’s mid-*p* correction to the Fisher’s exact test for AI analyses. After reviewing some of these methods, I will introduce a new practical significance test and compare the behavior of the 4/5ths rule, the Z_D test, the Z_{IR} test, and this newly developed test in various selection scenarios.

1.1 Adverse Impact

AI was discussed by the Supreme Court for the first time in *Griggs v. Duke Power Company* (1971). There were two important conclusions reached by the Supreme Court:

(a) the selection requirements disqualified Black applicants at a substantially higher rate than White applicants and (b) these requirements were not shown to be related with successful job performance. These conclusions indicate that it is not legal to use a selection requirement that is unrelated to job performance and, at the same time, disqualifies members of a racial/ethnic subgroup at a disproportionately higher rate. Seven years after the Supreme Court's conclusions, the *Uniform Guidelines* defined AI as: "substantially different rate of selection in hiring, promotion, or other employment decision which works to the disadvantage of members of race, sex, or ethnic group" **(**EEOC et al., 1978, Section 16B: Definitions, para. 3).

Thus, the most important part of the definition of AI given by *Uniform Guidelines* and of the conclusions reached by the Supreme Court is also the vaguest, regarding what constitutes a substantially different or higher rate. As stated in the *Uniform Guidelines*, the 4/5ths (or eighty percent) rule is the most accepted and well-recognized criterion in determining if the hiring or promotion rate is substantially different for various racial, ethnic or sex groups. Other than the 4/5ths rule, testing for a statistically significant difference between selection rates of race, ethnic, or sex groups is the other approach to determine if there is an evidence for AI (Collins & Morris, 2008; Morris & Lobsenz, 2000; Roth, Bobko, & Switzer III, 2006).

Although both the 4/5ths rule and the test of significance are considered as important informative tools to measure AI, the 4/5ths rule is often the primary reason an AI case is filed. Title VII of the 1964 Civil Right Act requires that a formal charge of discrimination needs to be processed by an agency such as EEOC (Landy, 2005). EEOC requires federal employees who want to file a discrimination case to contact an Equal

Employment Opportunity (EEO) Counselor at the agency where they work or applied for a job. If the dispute is not settled with the counselor, the next step is to file a formal complaint with the agency's EEO Office. If employees are not happy with the final decision by the agency's EEO Office, the next step is to appeal this decision to EEOC's Office of Federal Operations. If a settlement is not achieved during this process either, the final step is to file a lawsuit in the federal district courts. However, almost all of the cases are settled before they reach the final step. For example, there were 99,992 employment cases brought to EEOC's attention in the fiscal year 2010 (Charge Statistics, n.d.), while there were only 271 EEOC enforcement suits filed and resolved in the federal district courts (Litigation Statistics, n.d.). Knowing that federal agencies typically will only use the 4/5ths rule to detect AI (*Uniform Guidelines*, Questions & Answers, question 18), it seems that the 4/5ths rule is considered as the only benchmark in majority of the cases. After the cases reach the final step and brought to the court, however, the test of significance is often given more weight in the decision making process (Cohen & Dunleavy, 2009).

Next, I will discuss the current measurements of AI: the 4/5ths rule and significance testing.

1.2 The 4/5ths Rule

The 4/5ths rule states that if the selection rate for any minority groups falls under 4/5ths of the rate for the group with the highest rate; then, it is treated as a substantially different rate which, in turn, indicates evidence for AI.

In the areas of personnel selection research and employment law, the most common way of assessing the practical significance of AI is by use of the 4/5ths rule.

Bobko and Roth (2010) examined research articles in the applied psychological literature from 1990 to 2007 to determine which methods of assessment are typically used to identify if a selection procedure causes AI. The result of their study revealed that there were two typical ways: calculating a *d* value (standardized ethnic or gender group mean differences on the selection test) and comparing the hiring or passing rates of various subgroups. Six of the 24 articles used *d* values, and the remaining 18 articles attempted to detect AI by comparing hiring or passing rates. Among the articles comparing hiring or passing rate, 16 of them used the 4/5ths rule to assess AI while two of them just reported AI ratio without using the 4/5ths rule as a benchmark. None of the articles discussed significance testing on the difference between subgroup passing rates as a way of measuring AI. The result of this recent study clearly indicates that the 4/5ths rule is the most widely used and accepted measure of AI in the applied literature.

Despite its popularity, the 4/5ths rule is not flawless. One of the drawbacks of the wholesale application of the 4/5ths rule is its sensitivity not only to overall sample size, which affects the statistical power of any statistics, but also to both the total selection ratio (SR_T) and minority applicant ratio[5] (AR_{min}; Roth et al., 2006). More precisely, when SR_T and AR_{min} are smaller, the 4/5ths rule increases the probability of Type I error (i.e., finding AI when there is none). Conversely, when AR_{min} and SR_T are larger, there is an increase in Type II error rates (i.e., failing to indicate AI when it exists). Nevertheless, the effects of changes in AR_{min} and overall sample size on the behavior of the 4/5ths rule is

[5] Minority applicant ratio refers to the ratio of minorities in the applicant pool (*n* of minority applicants/ *n* of applicants) and is not to be confused with minority selection ratio (SR_{min}) which refers to the ratio of minorities who are selected to all minorities who are applied for the job (*n* of minorities selected/ *n* of minority applicants).

minimal, compared to the effect of changes in $SR_{T.}$ Thus, I will be focusing more on the effect of changes in SR_T on the behavior of the 4/5ths rule.

1.3 Statistical Significance Tests

Although significance testing is generally not a used practice for determining AI in applied literature (Bobko & Roth, 2010), the courts and regulatory agencies have recently begun to put more emphasis on tests of statistical significance (Cohen & Dunleavy, 2009). In line with this, there has been a recent increase in the number of simulations that have explored the use of various tests of significance as alternative or supplementary measurement to the 4/5ths rule. For example, Collins and Morris (2008), using simulations, discussed the changes in Type I error and statistical power rates for various tests of significance as a result of changes in AR_{min} and SR_T when the applicant pool size is small. These tests included the Z_D test, Fisher's exact test, Yates' continuity-corrected chi-square test, and a corrected chi-square test suggested by Upton (1982). Morris and Lobsenz (2000) compared Z_{IR} (the *Z*-test on impact ratio) and Z_D tests with regard to their statistical power and Morris (2001) compared the sample sizes required for the application of the Z_{IR} test, the Z_D test, and the 4/5ths rule in AI analysis. Roth et al. (2006) used Fisher’s exact test as a supplement to the 4/5ths rule. These studies provided different perspectives on the use of significance testing in AI assessment. However, the advantages and drawbacks of each type of significance test could benefit from further comparison.

The study by Collins and Morris (2008) concluded that the Fisher’s exact test and Yate’s chi square test were overly conservative if the applicant number was not exceptionally large and that these tests had lower power under many conditions. The

results for the corrected chi square test suggested by Upton (1982) and the Z_D test were comparable; both tests produced similar Type I error and power rates. The results further indicated that the Z_D test provided a better balance in maintaining a minimal Type I error rate while maximizing statistical power. Extrapolating on these findings, I decided to focus on the Z_D test along with the Z_{IR} test and excluded the other three tests from the analysis. The decision to use the Z_{IR} test along with the Z_D test was based on the conclusions by Morris and Lobsenz (2000). Their results indicated that the Z_{IR} test had a slight power advantage compared to the Z_D test. Therefore, I wanted to analyze how this slight power advantage influences the behavior of the Z_{IR} test relatively to the Z_D test.

Although I thought that it would be reasonable and informative to compare the behaviors of the Z_D and Z_{IR} statistical significance test along with the behaviors of the 4/5ths rule and the other practical significance test (proposed in this study), neither the *Uniform Guidelines* nor the OFCCP manual recommend any specific test of significance over other tests of significance. That means chi-square test, Fisher's exact test, the corrected chi square test suggested by Upton, and other tests of significance are all potentially legitimate options to assess AI. For example, the simulations by Collins and Morris (2008) demonstrated that the chi square test suggested by Upton (1982) is a reasonable alternative to the Z_D test when applicant pool size is small. Therefore, selection decision makers should closely follow the progress in the research area of AI measurement and use the test of significance recommended by the research for their specific selection context (i.e. small or large applicant pool size) to justify their decisions. Now, let us focus on the Z_D and Z_{IR} tests.

1.3.1 The Z_D Test

The Z_D test is conducted to measure if the difference between selection rates of majority and minority groups equals zero (Shoben, 1978). The proportions of minority and majority hiring are used as an estimation of the population selection rates for these groups. Subsequently, the difference between hiring rates is divided by its own standard error. If the absolute value of the result is larger than 1.96, the null hypothesis of H_0: $\pi_{min} = \pi_{maj}$[6] is rejected. The formula for the Z_D test is given below:

$$Z_D = \frac{SR_{min} - SR_{maj}}{\sqrt{SR_T(1 - SR_T)(\frac{1}{N_{min}} + \frac{1}{N_{maj}})}}$$

where SR_{min}, SR_{maj}, SR_T, N_{min}, and N_{maj} refer to minority selection ratio, majority selection ratio, total selection ratio, minority applicant number, and majority applicant number, respectively.

As Morris and Lobsenz (2000) pointed out, the Z_D test and the 4/5ths rule are based on different effect sizes. The effect size for the Z_D test is based on the difference between SR_{min} (minority selection ratio) and SR_{maj} (majority selection ratio), whereas the effect size for the 4/5ths rule is based on the impact ratio, the ratio of these selection ratios. This difference in the effect sizes constitutes a particular concern when there is a variation in SR_T, because the same IR will result in a smaller or larger difference in selection rates as SR_T changes. For example, a selection ratio difference of .05 will indicate an IR smaller than .80 when SR_T is small (e.g., the IR will be .50 when SR_{maj} and SR_{min} are equal to .10 and .05, respectively); however, the same difference will indicate an IR higher than .80 when SR_T is large (e.g., the IR will be .88 when SR_{maj} and SR_{min} are

[6] π_{min} refers to the population selection ratio for minority group, and π_{maj} refers to the population selection ratio for majority group. SR_{min} and SR_{maj} are estimators of π_{min} and π_{maj}, respectively.

equal to .80 and .75, respectively). Therefore, incorporating the results of the 4/5ths rule and the Z_D test conceptually becomes a difficult task. Arguing the difficulty of incorporating the results of the tests that are based on different effect sizes, Morris and Lobsenz (2000) introduced a significance test, the Z_{IR} test, which is based on the same impact ratio as the 4/5ths rule.

1.3.2 The Z_{IR} Test.

The Z_{IR} test is based on the IR and measures if the IR significantly different from 1.00. The sample IR (SR_{min}/SR_{maj}) is used as an estimation of the population IR (π_{min}/π_{maj}); then, natural logarithmic transformation of this estimated population IR is divided by its own standard error. If the absolute value of the results is higher than 1.96, the null hypothesis (H_o: $\pi_{min} = \pi_{maj}$) is rejected. The computational formula for the Z_{IR} test is presented below[7]:

$$Z_{IR} = \frac{\ln\left(\frac{SR_{min}}{SR_{maj}}\right)}{\sqrt{\frac{1-SR_T}{SR_T}\left(\frac{1}{N_{min}}+\frac{1}{N_{maj}}\right)}}$$

Although significance testing provided some insights on AI analysis, it is not without criticism. The main problem with significance tests is concerned with statistical power. As the research indicates, sample size has a tremendous influence on the power of statistical tests (Hsu, 1993). More precisely, a relatively large effect size might not reach statistical significance when the sample size is small; however, any nonzero effect size will reach statistical significance given a sufficiently large sample size. Therefore, the

[7] Interested readers are referred to Morris and Lobsenz (2000) for an extensive explanation of the mathematical and theoretical reasoning underlying the Z_{IR} test.

tests of significance are more likely to indicate a non-significant difference between SR_{min} and SR_{maj} as the applicant pool size get smaller and more likely to indicate a significant difference between SR_{min} and SR_{maj} as the applicant pool size gets larger. In addition to the applicant pool size, the magnitudes of SR_T and AR_{min} have a noticeable effect on the power of the tests of significance (Morris & Lobsenz, 2000); as the SR_T and AR_{min} approach .50, the power of the significance tests gradually increases, and a significant result becomes more likely. On the contrary, likelihood of obtaining a non-significant result increases as SR_T approaches either .00 or 1.00 and AR_{min} approaches .00.

Before discussing Type I and Type II error rates in the context of AI analysis, it is important to note the clear distinction between the 4/5ths rule as a test of practical significance and tests of statistical significance. Tests of statistical significance take degrees of freedom (sampling error variance) into consideration to determine if the results are significant. Thus, the same effect size obtained from 100 studies with varying sample sizes could result in 100 different *p* values (Thompson, 1999). That means the same selection ratio difference (SR_{maj}-SR_{min}) observed in 100 selection processes with varying applicant pool size could have 100 different *p* values. Some of these *p* values (for two-tailed test) could be smaller than .05, indicating evidence for AI; whereas others could be larger than .05, indicating no evidence for AI. However, note that a *p* value only gives the probability of observing a particular outcome (i.e., difference between SR_{min} and SR_{maj}) without specifying anything about the actual size of the effect. Accordingly, a trivial difference between SR_{min} and SR_{maj} may result in statistical significance with a *p* value smaller than .05 when applicant pool size is large (e.g., 5000), whereas a relatively

large difference between these ratios may result in a statistically non-significant *p* value larger than .05 when applicant pool size is small (e.g., 100).

Contrary to the tests of significance, the 4/5ths rule does not take the sampling error variance into account. The same effect size (impact ratio) is used for all conditions regardless of sample size. By an IR smaller than .80 indicating AI, the 4/5ths rule is intended to ensure that the effect is practically meaningful[8]; however, it fails to take the probability of obtaining a particular effect size into consideration. Previous simulation research on AI measurements clearly demonstrate that the probability of observing an IR smaller than .80 is typically substantially greater than the nominal alpha level of .05 when the applicant pool size and SR_T are small, whereas these probabilities are usually below the nominal alpha level when applicant pool size and SR_T are large (Greenberg, 1979; Roth et al., 2006). Therefore Type I error rates for the 4/5ths rule become a concern.

In short, each of the current AI measurements (tests of significance and the 4/5ths rule) has a particular problem: test of significance fails to take effect size into account, while the 4/5ths rule does not consider sampling error variance. It is a fairly easy task to compute effect size for a particular selection outcome. Therefore when the effect size satisfies the 4/5ths rule, but the significance test indicates AI, then selection decision makers can argue against the evidence of AI by bringing into attention that small effect sizes, even statistically significant ones, do not bear any practical implications. However, when the 4/5ths rule is violated by practically significant amount that indicates evidence for AI, selection decision makers can rely on the results of simulation studies to argue about whether sampling error variance might be an alternative explanation for this

[8] Although an impact ratio smaller than .80 could be considered as a meaningful effect size, the practical significance of this effect size gets smaller as the applicant number and selection ratio decrease.

finding. Simulation researchers run a lot of replications under the same selection conditions and obtain a different impact ratio for each replication. The distribution of these impact ratios provides selection decision makers with a base rate for situations where conclusions about AI are either consistent between the population and the sample (correct decision) or they are inconsistent (Type I and Type II errors).

1.4 Type I and Type II Errors in the Context of AI Assessment

The null hypothesis for the test of significance in AI analysis indicates that population selection ratio for minority group equals the population selection ratio for the majority group. When there is no subgroup mean differences on the selection test used and the variances of test scores are the same for minority and majority groups, a minority applicant and a majority applicant have the same probability of being hired if the locations of their test scores within their respective population test score distributions are the same. Put differently, the proportion of minority population that would perform well above a certain cutoff score will exactly be the same with the proportion of majority population that would perform well above this cutoff score. That is, in the presence of equal variance and absence of a subgroup mean difference on the selection measurement, the null hypothesis (H_0: $\pi_{min} = \pi_{maj}$) is always true. Therefore, finding a significant difference between selection ratios indicates Type I error. In fact, Type I error is the only worry in this context because when *d* equals zero in the population level, failing to reject the null hypothesis is not an error. Thus, Type II error does not pose any concern.

Contrariwise, when there is a meaningful subgroup mean difference on a selection measure, minority applicants and majority applicants have different probabilities of being hired even if the location of their test scores within their respective population test score

distributions are the same. That means proportions of minority and majority group populations that would perform well above a certain cutoff score on the selection measurement will be different. Consequently, rejecting the null hypothesis (H_0: $\pi_{min} = \pi_{maj}$) is the correct decision and failing to reject it constitutes an error (Type II error). As the population selection rates are different, rejecting the null hypothesis is not an error. Therefore, Type I error is not a concern in this situation.

Discussing Type I and Type II errors for the 4/5ths rule is not as easy as it is for the tests of significance. In a large part, this is because when discussing the 4/5ths rule, *Uniform Guidelines* considers the selection rates of minority and majority groups without specifying anything about the population selection rates. That is, the *Uniform Guidelines* does not consider the sample impact ratio as an estimate of the population impact ratio, and as an estimate, it must contain sampling error variance. In the context of the 4/5ths rule, sampling error variance should be considered as a factor when identifying whether or not the sample impact ratio correctly estimates the population impact ratio. However, this is not conceivable for the reason that the 4/5ths rule does not take sampling error into consideration, and that leads the 4/5ths rule to erroneously indicate AI when none exists in the population level (Boardman, 1979; Greenberg, 1979).

Thus, although there are clear conceptual differences between practical and statistical significance, there is no reason one cannot consider the statistical significance of statistics that indicate practical significance, such as the 4/5ths rule. Consider that 4/5ths rule produces Type I error whenever the IR is less than 4/5ths, indicating AI, yet the IR equal or greater than 4/5ths at the population level; conversely 4/5ths rule leads to a Type II error if it indicates no AI when there is in fact a meaningful subgroup mean

difference on the selection tests used. Although the investigation of Type I and Type II error rates for the 4/5ths rule is not directly in line with the language of the *Uniform Guidelines*, it is consistent with the spirit of those guidelines, because these error concepts serve to evaluate the accuracy of the 4/5s rule in a concrete manner.

It is relatively easy to talk about the Type I and Type II error rates at the extremes where subgroup mean difference are known to be nonexistent or very large, because then it is more obvious whether the 4/5ths rule will be met or violated. However, in typical situations, subgroup mean difference take on intermediate values (e.g., a *d* value of .30, .20 or lower), and conclusions about Type I and Type II error rates are unclear. It could be argued that the probabilities of observing AI should be lower in cases where *d* value is small ($d = .20$) and higher in cases where the *d* value is moderate or large ($d \geq .50$). These small, moderate, and large d values are based on the suggestion by Cohen (1992).

Using *d* values as benchmarks to evaluate Type I and Type II error rates of AI measurements would easily lead to a particular legitimate argument: Should selection decision makers simply rely on the *d*-value instead of the 4/5ths rule or the measurement proposed in this study? First, calculating a *d*-score reflecting the true population subgroup mean score difference might not either possible or feasible especially for small organizations dealing with small applicant pool and small selection ratios. For example, let us say that there are 150 job applicants and 18 of these applicants are minority. The mean score of these 18 minority applicants will not correctly reflect the true mean score of minority population. Therefore, these organizations might be required to conduct studies to estimate the true population subgroup mean difference on a test before using that test for selection purposes; but this will not be feasible for the organizations that

usually develop and use the selection test only once. Small organizations could not simply afford the time and money that is required to conduct such studies. Second, when it is feasible to conduct a study to estimate a *d* score, there is still an important final decision to be made: what benchmark to use to evaluate these *d* scores. For instance, does a *d* score of .23 indicate AI or not? Besides, the practical effect of a particular *d* score might not be the same for small and large sample sizes. For example, the practical effect of a *d* score of .15 will be negligible when the applicant pool size equals 200 and SR_T equals .20; however, the practical impact of the same *d* score will be enormous when applicant pool size equals 5000 and SR_T equals .50. Taking these points into consideration, it is clear that using *d* values as benchmark to evaluate AI is not straightforward as it seems to be.

1.5 Underlying Mathematical and Theoretical Reasons for the New Measure

Setting the null hypothesis that there is no subgroup mean difference ($d = .00$) between majority and minority groups on the selection measurement, the alternative hypothesis will be that there is a subgroup mean difference between these groups ($d \neq .00$). When, in fact, $d = .00$, previous research indicated that the Type I error rates for the 4/5ths rule are well above the nominal level of .05 if the selection ratio is small (e.g., .10, .20; Morris & Lobsenz, 2000; Roth et al., 2006). When the selection ratio approaches 1.00 (e.g., .80, .90), the 4/5ths rule indicates no AI in almost all cases. In other words, the Type I error rates really approaches zero[9]. There is no need to worry about Type II error in this scenario, because the null hypothesis ($d = .00$) is always true.

[9] The 4/5ths rule takes neither total selection ratio nor applicant pool size into account when measuring AI, but the tests of significance takes these variables into account. For this reason, the behaviors of these AI measurements can show quite different tendencies depending on the changes in total selection ratio and applicant pool size.

Why does the 4/5ths rule lead to high Type I error rate in cases where the selection ratios are small and lead no Type I error in almost all cases where the selection ratios are large? To answer this question, I computed the exact probabilities of obtaining a sample indicating AI for small and large total selection ratios. The first step was to calculate the possible number of samples one can observe from an applicant pool for a given selection ratio by using the combination formula:

$$\frac{n!}{h! * (n-h)!}$$

Here, *n* is the total number of applicants and *h* is the number of applicants hired. Having a selection ratio of .24 and a total of 100 applicants, I get:

$$\frac{100!}{(100-24)! * (24)!} = 79{,}776{,}075{,}565{,}900{,}400{,}000{,}000$$

The outcome above is the exact number of possible samples that consist of 24 hires from a total of 100 applicants.

As the second step, I computed the minimum required number of minority hires which does not violate the 4/5ths rule for a given minority selection ratio, such as .20. Based on the total applicant number of 100, SR_T of .24, and AR_{min} of .20[10], the minimum number of minority applicants to be hired should be at least 4 not to violate the 4/5ths rule. When there are 4 minorities hired, SR_{min} will be .20 $(4/20)$ and SR_{maj} will be .25 $(20/80)$ resulting in an IR of $.20/.25$ or 4/5, which indicates no AI. Therefore, the next step is to compute the possible number of samples that consist of 0, 1, 2, or 3 minority applicants. The number of these possible samples was computed again by using

[10] 100 for applicant pool size, .20 for minority ratio, and .24 for selection ratio were selected for ease of computation.

combination formula. For example, the number of possible samples with no minority applicant equals:

$$\frac{(N_{maj})\,!}{(N_{maj}-h)!\,(h)!}$$

or,

$$\frac{80!}{(80-24)!*(24)!} = 162{,}238{,}272{,}822{,}100{,}000{,}000$$

And, the number of possible samples with only one minority applicant equals:

$$\frac{80!}{(80-23)!*(23)!} * \frac{20!}{(20-1)!*(1)!} = 1{,}366{,}217{,}034{,}291{,}370{,}000{,}000$$

The number of possible samples with exactly two and three minority hires was computed in the same way. Summing these numbers provides us with the exact number of possible samples with 0, 1, 2, or 3 minority hires, which equals to 18,190,354,372,162,800,000,000. This sum also equals to the number of possible samples violating the 4/5ths rule. Next, dividing this sum by the number of possible sample with 24 hires (which equals 79,776,075,565,900,400,000,000), I get .228017663730981; this is the exact probability of obtaining a sample indicating AI for the case where n = 100, SR_T = .24, and AR_{min} = .20. The same procedure was followed to compute the exact probabilities of observing AI for the selection ratios of .12 and its multiples and AR_{min} of .10 and .20[11]. The results of these computations, presented in Figure 1, give the probabilities of observing AI when completely random sampling method used in hiring. These exact probabilities for the same SR_T and AR_{min} will change as the applicant pool size changes. Although a decrease in these exact probabilities is expected, in general, as the applicant pool size increases, sometimes a particular increase in applicant pool size

[11] The values of .12 and its multipliers for selection ratio and .10 and .20 for minority applicant ratio are selected for ease of computation.

can lead to a dramatic increase in the exact probabilities of observing AI for a particular SR_T and AR_{min} combinations. This is particularly true when SR_T, AR_{min} and applicant pool size are small. For example, when applicant pool size, SR_T, and AR_{min} respectively equal 100, .12, and .10, the probabilities of observing AI is .26075 (see Figure 1), when we change only the applicant pool size from 100 to 110, however, the probabilities of observing AI increases to .61679[12]. Therefore, when examining Figure 1, the reader should always keep in mind that these exact probabilities are only true for the given conditions and that the main purpose of this figure is to illustrate the trend in the exact probabilities of observing AI as a response to the changes in SR_T.

Another important point to note is that when $d = 0$, there is no AI in the population level. Therefore, AI probabilities presented in Figure 1.1 are due to the sample size being less than infinity. That means, if it was possible to run a simulation with an infinite number of applicants, the results would perfectly reflect the population parameters (i.e., indicating no AI when $d = 0$ in the population level). Understanding this is crucial for a complete understanding of the influence of sample size on AI when $d = 0$. In short, smaller sample sizes produce a larger negative gradient, whereas larger sample sizes produce smaller negative gradient in response to changes in SR_T. The trend will be the same as the sample size approaches infinity. More precisely, the gradient will become lower and lower, and will be zero eventually.

[12] Observing such an increase in the exact probabilities of observing AI when there is a small change in the applicant pool size is due to fact that when SR_T and applicant pool size are small hiring or not hiring one more minority applicant substantially affects the magnitude of impact ratio.

Figure 1.1 The Probabilities of Observing AI for the 4/5ths Rule when $d = .00$ and $n = 100$.

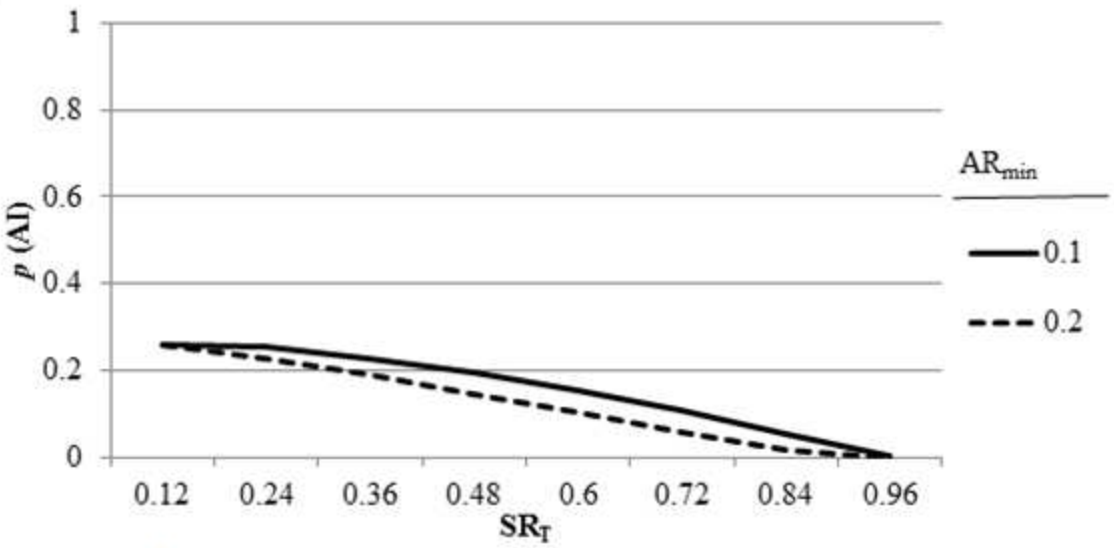


Figure 1.1 These probabilities are only valid for the given SR_T, AR_{min}, and applicant pool size combination. Here applicant pool size is set to be 100, SR_T is set to range from .12 to .96, and AR_{min} is set to.10 and .20.

Figure 1.1 illustrates the changes in the probabilities of observing AI as a response to changes in selection ratios where a complete random sampling method is employed, the applicant pool size is set to 100, and the minority base rate is set to 10% or 20%. Results indicated that more than 20 percent of the time the 4/5ths rule wrongly concluded that the selection procedure causes AI in cases where the total selection ratio was .36 or lower. For example, when AR_{min} equals .10 and SR_T equals .24, Type I error rate or the probability of observing AI is .25. In other words, applying the 4/5ths rule as a standard of practical significance leads to higher Type I error rates than the nominal level of .05. Taking these high Type I error rates into consideration, the following proposition is offered:

1.6 Research Proposition 1

Improved alternatives to the 4/5ths rule as a standard for practical significance should behave more conservatively by decreasing the probabilities of observing AI in cases where subgroup selection ratios are small.

The exact probabilities of observing AI for the 4/5ths rule decreased below the nominal level of .05 when the selection ratios were .84 or higher. In addition, the results revealed that the probabilities of observing AI were around .003 and .000, respectively for the minority subgroup base rate proportion (AR_{min}) of .10 and .20, when SR_T was .96. Having an exact probability of zero implies that the 4/5ths rule will indicate no AI regardless of the magnitude of subgroup mean difference when SR_T is large enough. Although the utility of using a selection tool where the SR_T equals to .96 can be argued against in most cases, it clearly illustrates the possibility that the 4/5ths rule might lead to Type II error when SR_T is large. That is, with a high SR_T the 4/5ths rule might not indicate AI, when, in fact, there is a meaningful subgroup mean difference. For that reason, the proposition follows:

1.7. Research Proposition 2

Relative to the 4/5ths rule, any alternative rule should ideally behave in a less conservative way by increasing the probabilities of observing AI when the SR_T is large.

To address the two propositions above, I computed an alternative IR by using logarithmically transformed minority and majority selection ratios instead of raw selection ratios. This new IR will be referred as IR_{Ln} from now on through the paper, and it equals:

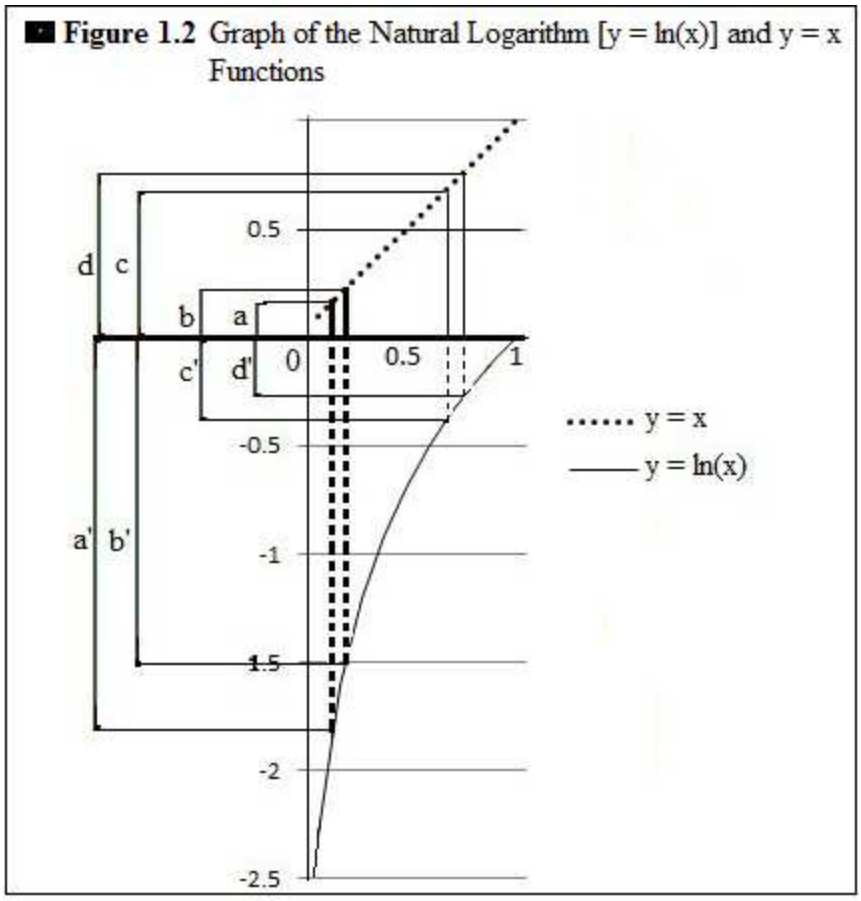
$$IR_{Ln} = Ln(SR_{maj}) / Ln(SR_{min})$$

Adopting the same procedure with the 4/5ths rule but using IR_{Ln} instead of IR, an alternative AI measurement was developed. This measurement will be referred as the Ln rule, and it is adopted because, in line with the previous two research propositions, it behaves more conservatively than the 4/5ths rule when the selection ratios are small and less conservatively than the 4/5ths rule when the selection ratios are high. It is important to note explicitly that the Ln rule is *not* the same as $\ln(SR_{maj}/SR_{min})$, which serves the different goal of transforming a wide range of impact ratios to a normal distribution.

The reason behind using logarithmically transformed ratios instead of raw ratios to develop a rule that is more conservative than the 4/5ths rule when SR_T is small and less conservative than the 4/5ths rule when SR_T is large can be explained concretely by referring Figure 1.2. $Y = x$ and $y = \ln(x)$ functions in Figure 1.2 will be referred when discussing the behaviors of the 4/5ths rule and the Ln rule, respectively. *Y*-axis values in $y = x$ function indicate the possible values that the numerator and denominator of the 4/5ths rule's impact ratio can get. Similarly, *y*-axis values in $y = \ln(x)$ function indicate the possible values that the numerator and denominator of the Ln rule's impact ratio can get. There are three important differences between $y = x$ and $y = \ln(x)$ functions. First, *y*-axis values range from 0 to 1.0 for $y = x$ function while those values ranges from 0 to negative infinity for $y = \ln(x)$ function. Second, the slope of $y = x$ function remains stable and always equals 1.0 while the slope of $y = \ln(x)$ function equals 1.0 when x equals 1.0 and increase hyperbolically as x approaches 0. Third, *y*-axis values get smaller as x approaches 0 for $y = x$ function whereas *y*-axis values get bigger in absolute value as x approaches 0 for $y = \ln(x)$ function. These differences are the key factors shaping the behavior of the 4/5ths rule and the Ln rule when assessing AI.

When discussing the effect of SR_T on the behavior of the 4/5ths rule and the Ln rule, I will assume that both minority and majority selection ratios will be around SR_T; therefore, a small SR_T means small SR_{min} and small SR_{maj}. Although this assumption is not always true[13], it is true most of the time. Furthermore, the main concern in AI assessment is to correctly identify if there is an evidence for AI when SR_{min} and SR_{maj} do not differ very much from each other (i.e., it is more likely to be accurate to find evidence for AI when there is a large difference between SR_{min} and SR_{maj}). If, for example, SR_{maj} is meaningfully larger than SR_T, then SR_{min} must be meaningfully smaller than SR_T. When this is the case, both the 4/5ths rule and the Ln rule will correctly indicate AI. Thus, assuming that both SR_{min} and SR_{maj} will be around SR_T does not pose any concern when describing the logic behind the use of logarithmically transformed ratios.

[13] The possible values SR_{min} and SR_{maj} can get depend not only on SR_T but also on minority and majority applicant ratio. For example, the possible minimum and maximum values of SR_{maj} are .29 and .71, respectively, when SR_T = .5 and AR_{maj} = .7. When AR_{maj} decreases to .55, the possible minimum and maximum values of SR_{maj} become .09 and .91, respectively.



Now, let us focus on the left part of these functions where x is equal to or smaller than .50 to compare the behavior of the 4/5ths rule and the Ln rule when SR_T is small. Remember that Type I error is the main concern for the 4/5ths rule when SR_T is small. In other words, the 4/5ths rule is more likely to erroneously indicate higher rates of AI in cases where SR_T is small. This is because, holding the difference between SR_{min} and SR_{maj} constant, the ratio (IR) of these two ratios gets smaller and smaller as they approach zero. Therefore, the 4/5ths rule, relying on these raw ratios, becomes more likely to indicate AI when there is actually a small and impractical selection ratio difference that would be expected due to chance. Using the raw ratios, for example, the ratio of *a* to *b* (.17/.22) is .77 (see Figure 1.2) which is below the benchmark of .80. That means a selection ratio difference of .05 will indicate AI. Let us focus on smaller

selection ratios to make the argument clearer. Using raw ratios of .07 and .09, the IR is .78 and it indicates evidence for AI despite the fact that selection ratio difference is as small as .02. This suggests that a less conservative rule is more appropriate to measure AI when selection ratios are small. To comply with this suggestion, I used natural logarithmically transformed ratios instead of raw ratios to compute an IR that is directly comparable to the 4/5ths rule's benchmark of .80. The IRs computed using natural logarithmically transformed ratios are always larger than the IRs computed using raw ratios, so long as both SR_{min} and SR_{maj} are lower than .38. Referring to Figure 1.2, the ratio[14] of ln(*b*) to ln(*a*) (or *b'/a'*) equals -1.51/-1.77 or .85, which is larger than the 4/5ths rule's benchmark of .80. Using smaller ratios, the ratio of ln(.09) to ln(.07) is .91, which is not only larger than the IR obtained using raw ratios but also larger than the benchmark of .80. As this example illustrates, the Ln rule behaves in a more conservative manner than the 4/5ths rule when SR_T is small. By requiring a larger difference between SR_{min} and SR_{maj}, the Ln rule decreases Type I error rates when the selection ratios are low.

There are two important characteristics of $y = \ln(x)$ function that help the Ln rule behave more conservatively than the 4/5ths rule when SR_T is small. First, *y*-axis values in $y = \ln(x)$ function are larger in absolute value than y-axis values in $y = x$ function for small selection ratios. Second, the slope of $y = \ln(x)$ is larger than the slope of $y = x$, and it gets larger as x approaches zero. Recall that *y*-axis values in $y = x$ represent the array of possible values of SR_{min} and SR_{maj} for the 4/5ths rule's, and *y*-axis values in $y = \ln(x)$ represent the array of possible values of the numerator and denominator of the Ln rule's

[14] The natural logarithmic-transformed ratio of a raw ratio increases in absolute value as that raw ratio decreases. That means a lower minority ratio will be larger, in absolute value, than a higher majority ratio. Therefore, I divided $\ln(SR_{maj})$ by $\ln(SR_{min})$ to get the new IR that is directly comparable to the 4/5ths rule's IR.

IR. Focusing on the left part of y = ln(x) function, a natural logarithmically transformed ratio is larger than the corresponding raw ratio in a way such that it is impossible to obtain an IR that is lower than .80 if there is not a large difference between two naturally log transformed ratios. The steep slope of y = ln(x) function makes it possible to observe a large difference between two natural log transformed ratios. Therefore, observing IRs smaller than .80 become possible for the Ln rule when SR_T is small. In short, the steep slope of y = ln(x) compensates the effect of having large (in absolute value) numerator and denominator when computing an IR for the Ln rule.

Now, let us focus on the right part of $y = x$ and y = ln(x) functions where x is greater than .50 to compare the behavior of the 4/5ths rule and the Ln rule when SR_T is large. When SR_T is large, as discussed earlier, the 4/5ths rule fails to indicate AI in cases where minority and majority selection ratio difference as large as .15 is evident. Put simply, the 4/5ths rule is more likely to erroneously indicate no AI (Type II error) when SR_T is large. The reason for that lays within the fact that IR gets larger and larger as SR_{min} and SR_{maj} approach 1.00. This is always true as long as the difference between SR_{min} and SR_{maj} is kept constant and SR_{min} is smaller than SR_{maj}. Relying on these raw ratios, the 4/5ths rule becomes less likely to indicate AI when SR_T is large regardless of how practical the selection ratio difference is. Using the raw ratios, for example, the ratio of c to d (.65/.78) is .83, which is above the 4/5ths rule's benchmark of .80. Note that the difference between SR_{min} and SR_{maj} is .13. However, using natural log transformed ratios, the ratio of ln(d) to ln(c) or d'/c' (-.25/-.43) is .58, a value below .80. Let us consider larger selection ratios to make the discussion clearer. Using raw ratios, the ratio of SR_{min} of .76 to SR_{maj} of .95 is .80. That means a selection ratio difference of .19 will not

indicate AI when the 4/5ths rule is used as a benchmark. Using naturally log transformed ratios, on the other hand, the IR_{Ln} will be .19 [ln(.95/ln(.76)], indicating evidence for AI. Therefore, behaving less conservatively than the 4/5ths rule as SR_T approaches 1.00, the Ln rule reduces possible Type II error rates in cases where a large subgroup mean difference is evident.

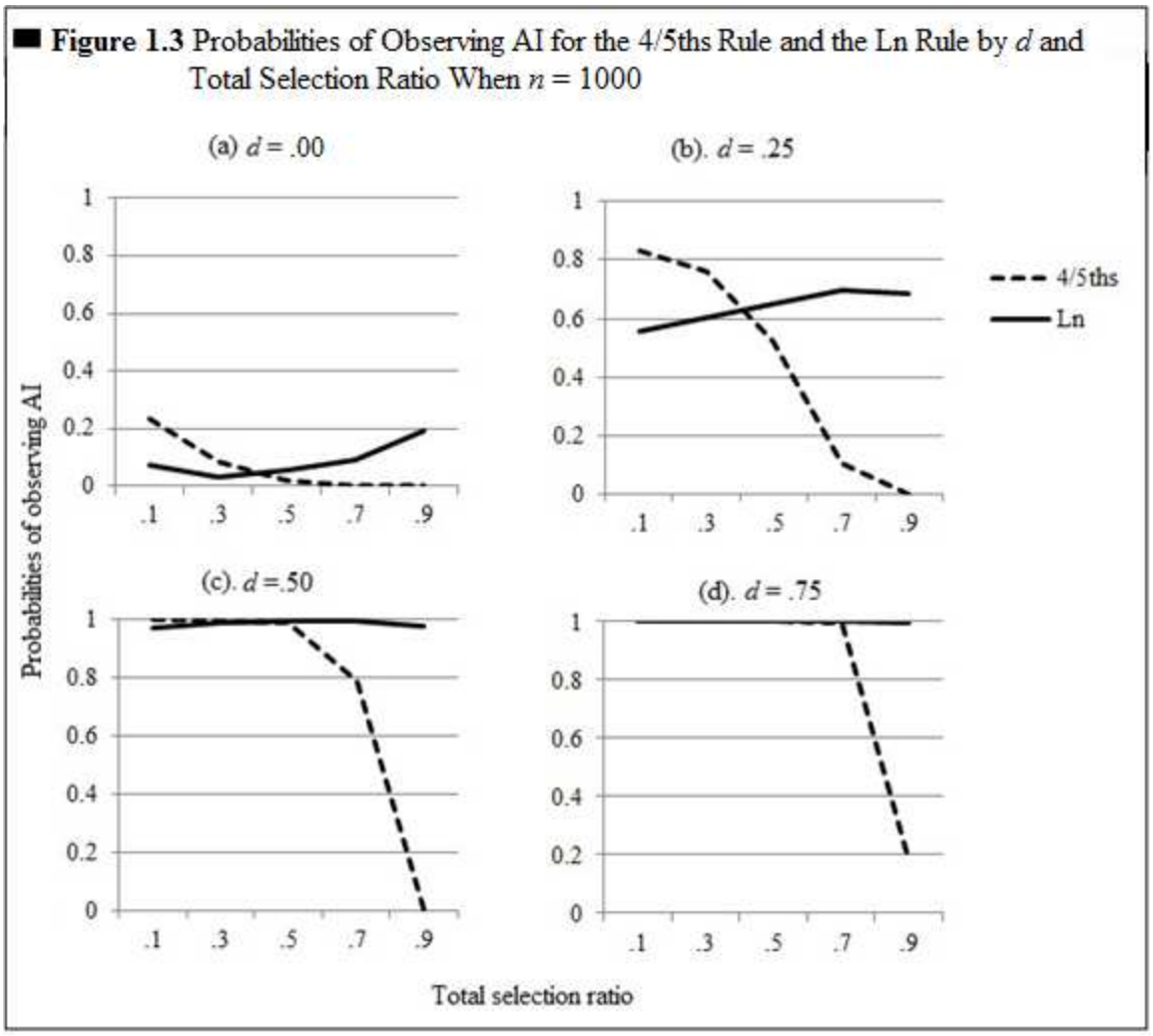
The main reason why the Ln rule behaves more conservatively than the 4/5ths rule when SR_T is large can be understood, again, by comparing $y = x$ and $y = \ln(x)$ functions. *Y*-axis values in $y = x$ function get larger as x approaches 1.00 while y-axis values in $y = \ln(x)$ function get smaller (in absolute value) as x approaches 1.00. Remember that here x represents the array of possible values that numerator and denominator of 4/5ths rule's and the Ln rule's IR can get. As previously mentioned, there is a linear relationship between an IR and numerator and denominator (which represent SR_{min} and SR_{maj}, respectively) of that impact ratio if numerator is smaller than denominator. That means IR gets larger as the numerator and denominator get larger and IR gets smaller as these values gets smaller as long as the difference between numerator and denominator is kept constant. As SR_T increases, numerator and denominator of the 4/5ths rule's IR get larger in a way that the 4/5ths rule fails to identify a significant and practical difference between numerator (SR_{min}) and denominator (SR_{maj}) as an evidence for AI

Focusing on the right part of $y = \ln(x)$ function where x is greater than .57, natural logarithmically transformed ratios are smaller than the corresponding raw ratios. Using these transformed ratios (numerators or denominators in this case), the Ln rule compared to the 4/5ths rule leads to lower IR. Therefore, the Ln rule behaves less conservatively

than the 4/5ths rule when SR_T is large. That means the Ln rule is more likely to signal the presence of evidence of AI when there is a practical and significant difference between SR_{min} and SR_{maj}. In other words, the Ln rule compared to the 4/5ths rule reduces Type II error rates when SR_T is large.

The reduction in Type I error rates when SR_T is small and Type II error rates when SR_T is large is, however, not a gain without cost. First, an increase in Type II error rates may become inevitable when a less conservative measure is used in cases where SR_T is small. Second, adopting a more conservative measure when the selection ratio is large may lead to an increase in Type I error rates. Therefore, the Ln rule should preserve a better balance than the 4/5ths rule between Type I and Type II error rates, thereby becoming a statistic that incorporates both statistical and practical significance. The 4/5ths rule is an index of practical significance that ignores all differences in selection conditions, and it does not consider statistical significance whatsoever. The formulation of the Ln rule addresses both of these critical issues.

To explore if the Ln rule is truly successful in providing this balance, I conducted a set of simulations and compared the behavior of the Ln rule to the behavior of the 4/5ths rule in situations where the subgroup mean difference (*d* value) was set to .00, .25, .50, or .75. The same procedure described in the Method section was also followed for these simulations, and the AR_{min} and total applicant pool size were set to be .12 and 1000, respectively. The results are presented in Figure 1.3.



Contrary to expectations, the simulations indicated that if there was a medium or large subgroup mean difference (a *d* value of .50 or .75) on the selection measure, Type II error rates for the Ln rule and for the 4/5ths rule were almost the same when the total selection ratio was equal or smaller than .50. It means the decrease observed in Type I error rates when there was no or small subgroup mean difference was obtained without a noticeable cost.

In some cases, organizations might not successfully attract as many applicants as possible in order to be selective in their hiring decisions. Regardless of whether this was due to an unsuccessful recruitment campaign or a shortage in the workforce,

organizations might end up with a need to hire a large percentage of the applicants in the applicant pool. In such a situation, a SR_T of .80, .90 or even .95 is very likely. Bobko and Roth (2004) discussed a court case where the plaintiff argued that using the ratios of rejection rates instead of hiring rates signaled that the selection procedure caused AI against minority applicants. The SR_{min} was .96 and SR_{maj} was .98; the rejection rates were .04 and .02, respectively for minority and majority applicants and thus the rejection ratio was .02/.04 = .50. The plaintiff claimed that because the rejection ratio was below 4/5, it indicated evidence for AI. Nevertheless, the total selection ratio was at least .97 in this situation.

As presented in Figure 1.1, using the 4/5ths rule as a benchmark, the exact probability of obtaining a sample indicating AI is zero when the total selection ratio is .96 and AR_{min} is .20. That means, regardless of how large the subgroup mean difference on the selection test is, the 4/5ths rule will fail to identify AI against minority applicants when the total selection ratio is large enough. How large a total selection ratio needs to be to observe this situation, however, directly depends on the minority applicant ratio. As the minority applicant ratio increases, the total selection ratio needed to observe this situation decreases. For example, when minority applicant ratio is .50, a total selection ratio of .90 will be a value large enough to observe this effect. Furthermore, previous research concluded that the probability of observing AI decreased for the 4/5ths rule, as SR_T increased (Bobko & Roth 2004). This decrease went down to the acceptable alpha level of .05, as the SR_T approached 1.00. If there was small subgroup mean difference (*d* = .20 or .30) or no difference at all (d = .00), this decrease in the probabilities of observing AI is desirable because it indicates a decrease in Type I error rate. But, what is the effect of

this decrease on Type II error rates? As discussed before, Type II error rate was a concern only when there is a significant standardized mean difference between minority and majority groups.

The simulation results presented in graphs (c) and (d) of Figure 1.3 indicated that when SR_T was as large as .90, the 4/5ths rule indicated AI less than 20 percent of the time even in the situations where a medium or a large subgroup mean difference was evident. Behaving in a less conservative manner when total selection ratios were large, the Ln rule was more likely to indicate AI. That is, the Ln rule outperformed the 4/5ths rule when there was a medium or large subgroup mean difference.

When *d* value was small (*d* = .00 or .25), on the other hand, the behavior of the Ln rule was more problematic than the behavior of the 4/5ths rule in situations where SR_T was larger than about .40. As the graphs (a) and (b) of Figure 1.3 demonstrated, the Ln rule was more likely to indicate AI than the 4/5ths rule as SR_T approached 1.00. For example, when *d* = .25 and SR_T = .70, the Ln rule and the 4/5ths rule indicated AI 69% and 10% of the time, respectively. That means, as SR_T approached 1.00, Type I error rates for the Ln rule increased well above the Type I error rates for the 4/5ths rule. To mitigate this problem, I made some adjustment in the Ln rule by taking the effect of the changes in SR_T into account. The adjustment was to multiply the IR for the Ln rule (IR_{Ln}) by $\ln(.8*SR_T+2.63)$. When SR_T was around .10, the value of this expression would be around 1.00; as SR_T increased, the value of this expression would increase above 1.00. Multiplying IR_{Ln} by a value larger than 1.00, the resulting value (or the adjusted IR_{Ln}) will be higher than the actual IR_{Ln}. Therefore, as SR_T approached 1.00, the adjusted Ln rule (Ln_{adj}), relative to the Ln rule, was less likely to indicate AI. Being less likely to

indicate AI in cases where SR_T is large, Ln_{adj} will produce less Type I error than the Ln Rule. Ln_{adj} is given below:

$$\mathrm{Ln}_{adj} = \frac{\mathrm{Ln}(\mathrm{SR}_{maj})}{\mathrm{Ln}(\mathrm{SR}_{min})} * \mathrm{Ln}(0.8 * SR_T + 2.63)$$

When making the adjustment, there were two concerns. First, the adjusted rule needed to be less sensitive to the changes in SR_T and applicant pool size and more sensitive to changes in *d* values on the predictor. Second, knowing that the Ln rule behaves less and less conservatively as total selection ratio approaches 1.00 and thus increases Type I error rates when there is small or no subgroup mean difference, the adjusted rule should be more conservative (i.e., less likely to indicate AI) than the Ln rule as the total selection ratio increases. I made various adjustments to the Ln rule to address these concerns. Some of these adjustment decreased Type I error rates when SR_T was small but increased Type II error rates when SR_T was large while others decreased Type I error rates when SR_T was large but increased Type II error rates when SR_T was small. The proposed adjustment, although not being perfect, did relatively well comparing to the other adjustments in compensating both Type I and Type II error rates.

There is not a pure or clear mathematical explanation for the adjustment made. I used an applied method of successive approximation to adjust the Ln rule in a way that the new rule provides a balance between Type I and Type II error rates. First, I set n = 400 and 1000, AR_{min} = .12 and *d* = .25 to create two base conditions to test the effect of various iterations of adjustment on the behavior of Ln_{adj}. These base conditions were deliberate choices because AR_{min} of .12 best represents the proportion of Black and Hispanics in the workforce (U.S. Department of Labor & U.S. Bureau of Labor Statistics, 2011), applicant sample of 400 and 1000 indicate moderate sample sizes (compared to a

smaller sample size of 200 or a larger sample size of 5000 used in the simulations), and a *d* value of .25 ensures observing the effect of adjustment across all selection ratios[15]. Although relying on the method of successive approximation rather than a theoretically solid approach to adjust the Ln rule seems ineffective in nature, it is indeed the strongest part of Ln_{adj}. Previous research has repeatedly showed the above presented deficiencies of the 4/5ths rule. The method of successive approximation made it possible to minimize these deficiencies trial after trial. Figure 1.4 provides clear insight into how these trials are successful in finalizing the adjustment made.

Figure 4 provides two distinct information structures. First, the functions of .05 SR diff (ARmin= .12) and .05 SR diff (ARmin= .20) show the changes in the probabilities of observing a .05 selection ratio difference between minority and majority subgroup as a response to changes in SR_T respectively for AR_{min} of .12 and .20 in condition where *n* equals 1000 and *d* equals .00[16]. Second, the other three functions shows the changes in the maximum values selection ratio difference between minority and majority subgroups can get to comply with the corresponding AI measurements as a response to changes in SR_T. Selection ratio differences larger than the values given in these second set of functions violate the corresponding AI measurement.

[15] Most of the adjusted Ln rules correctly indicate AI almost in all cases when *d* equals or larger than .50 or correctly indicate no AI almost in all cases when *d* equals .00. Therefore, it was not possible to observe any meaningful variations between subsequent iterations of adjustment in these conditions where none or moderate to high subgroup mean difference was evident.

[16] I run simulations to obtain these probabilities. The same method described in the method section was also used in here.

Figure 1.4 Probabilities of Obtaining a .05 SR Difference When $n = 1000$, $AR_{min} = .12$ or .20, and $d = .00$; and Maximum SR Difference to Comply with the 4/5ths, Ln_{adj}, and Ln Rule by SR_T.

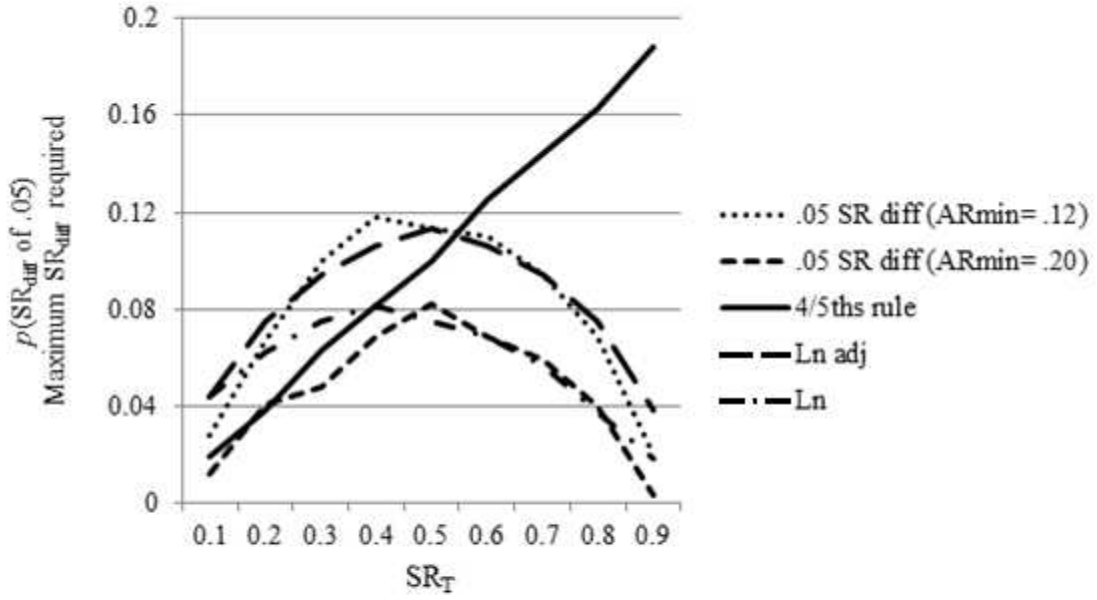


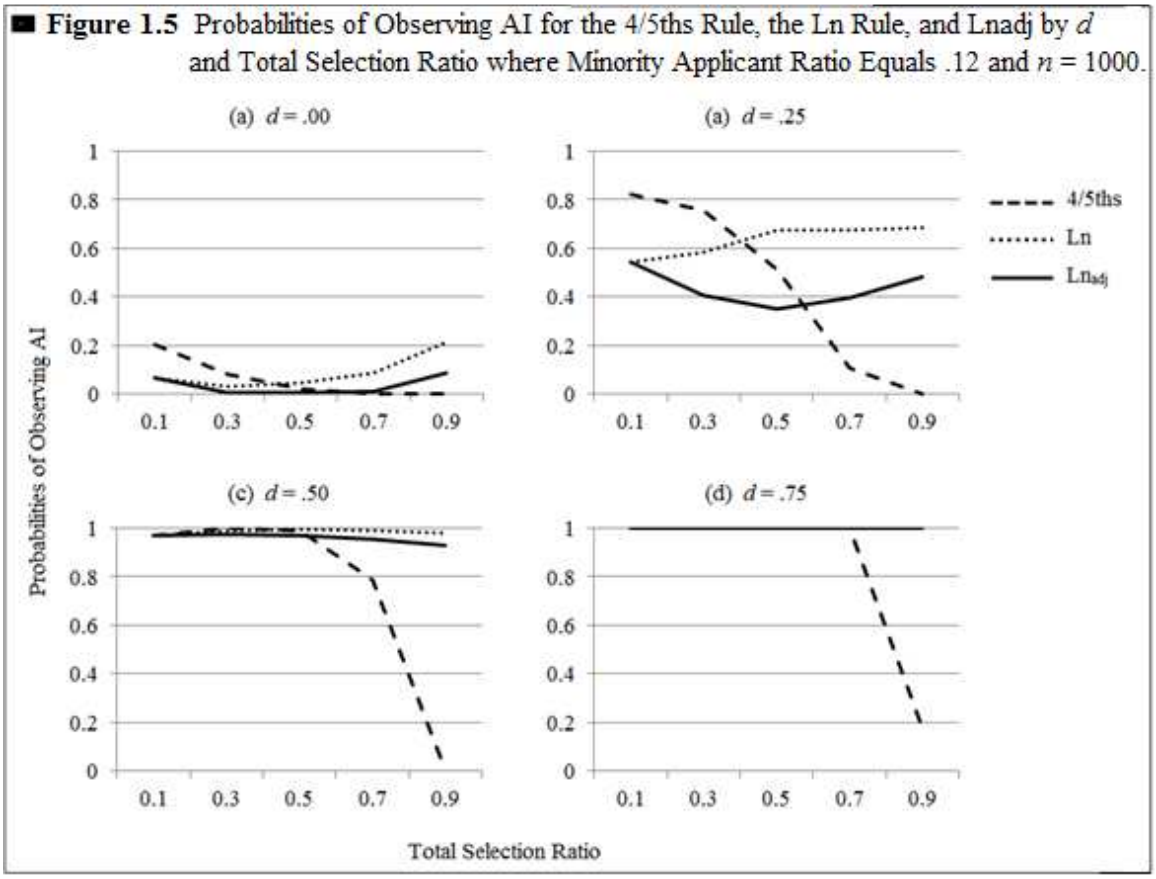
Figure 1.4 This figure provides two distinct information structures. The lines of .05 SR difference show the changes in the probabilities of observing a .05 selection ratio difference (between minority and majority selection rates) as a result of change in SR_T (total selection ratio) when $n = 1000$, $AR_{maj} = .12$ or .20, and $d = .00$. The other lines show the changes in the maximum value SR difference can get to comply with the corresponding AI measurements as a response to changes in SR_T

Figure 4 provides inverted-U shaped functions for probabilities of observing a particular selection ratio difference (e.g., a SR difference of .05 in this case). That means, the probabilities of observing a particular SR difference are small for either small or large SR_T and large for moderate SR_T. The shape of Ln_{adj} function matches best to the shape of .05 SR diff functions. This indicates that Ln_{adj} requires a small selection ratio difference to indicate AI when the probabilities of observing a particular difference is small and requires a large selection ratio difference when the probabilities of observing a particular difference is large. The function of the Ln rule is also provides a good match with SR difference functions, but it has positive skewness. The function of the 4/5ths rule, on the

other hand, does not show even a slight match with SR diff functions. This is because the 4/5ths rule does not consider the changes in the probabilities of observing a particular effect size as a response to changes in SR_T when measuring AI. Although the probabilities of observing the same effect size is very similar when SR_T is either small or large, the 4/5ths rule requires a very small selection ratio difference when SR_T is small and requires a very large selection ratio difference when SR_T is large. Figure 1.4, in this regard, shows that $Ln_{adj,}$ compared to the 4/5ths and the Ln rule, is an efficacious measurement that considers the changes in the probabilities of observing a particular effect size into account and behaves accordingly. This also clearly demonstrates that although being practical significance tests, Ln_{adj} and the Ln rule take probability theory into consideration. Examining the next figure (Figure 5) will further help the reader to understand how the adjustment made changed the behavior of the Ln rule.

Figure 5 presents the results of simulations comparing the behavior of the 4/5ths rule, the Ln rule, and Ln_{adj}. As presented in graphs (a) and (b) of Figure 1.5, the behavior of Ln_{adj} is different from the behavior of the Ln rule when there is a small (d = .25) or no subgroup mean difference (d = .00). As predicted, Ln_{adj} is less likely to indicate AI than the Ln rule as SR_T increases above .10. Remember that Type I error rates are the main concern when there is small or no subgroup mean difference. Therefore, being less likely to indicate AI than the Ln rule in these cases, Ln_{adj} further decreases Type I error rates. Obtaining such a decrease is particularly important in cases where SR_T equals or larger than .50 because the Ln rule, behaving less conservatively than the 4/5ths rule, increases the probability of Type I error in those cases. Thus, a decrease observed in the probabilities of obtaining AI as a result of the adjustment when there is a small or no

subgroup mean difference is a desired outcome. Although this decrease indicates less Type I error rates when d = .00 or d = .25, it has a potential to cause an increase in Type II error rates when d = .50 or d = .75. Examining the graphs (c) and (d) of Figure 1.5, the results indicated that when there is a moderate or large d value, the behavior of the Ln rule and Ln_{adj} are almost the same. That means the decrease in the Type I error rates observed in graph (a) and (b) of Figure 1.5 is obtained without a noticeable cost.



In short, decreasing Type I error rates without increasing Type II error rates, the results for Ln_{adj} were more promising. For the reason being that, I measured the behavior of Ln_{adj} and compared it to the behaviors of the 4/5ths rule, Z_{IR} test and Z_D test.

Chapter 2

Method

The purpose of this study was to make comparisons among the behaviors of the 4/5ths rule, Ln_{adj}, the Z_D test, and the Z_{IR} test when assessing AI under a variety of conditions. These conditions include variation in (a) total selection ratio (SR_T), (b) applicant pool size (n), (c) minority applicant ratio (AR_{min}), and (d) the subgroup mean difference (d). First, I simulated the data assuming that there was no subgroup mean difference (d = .00) on the selection test used. Then, the data were simulated by assuming that there were small (d = .25), medium (d = .50), or large (d = .75 and 1.00) subgroup mean differences on the selection test. Simulating data with no or small subgroup mean difference allowed observing Type I error rates for each of the four AI measurements. Simulating data with a medium or large subgroup mean difference, conversely, allowed comparing the behavior of these AI measurements in conditions where Type II error was a concern.

2.1 Simulating Data with no Subgroup Mean Difference

I used R code version 2.12.2 to simulate the data. First, a majority applicant pool was generated by selecting majority applicants randomly from a simulated population with a mean of 100 and standard deviation (SD) of 15. I set these values because IQ is scaled to a mean of 100 and a SD of 15 in the population level (Neisser et al., 1997) and I wanted to use a well-known distribution instead of creating a normal distribution by assigning an arbitrary mean and SD. Then, minority applicant pool was generated by selecting minority applicants randomly from a simulated population with the same mean

and SD. Second, these two applicant pools were combined into one overall applicant pool. After that, a top-down selection procedure was used and applicants were selected until there was no open position left for a particular SR_T and applicant number (*n*) combination. This procedure was repeated 1000 times for each combination of AR_{min}, SR_T, and *n*. Based on the minority selection ratio (SR_{min}) and majority selection ratio (SR_{maj}) obtained from each repetition, I computed (a) the average difference (SR_{diff}) between SR_{maj} and SR_{min}, (b) the average sample impact ratio ($IR_{average}$), (c) the number of repetitions that were resulted in violation of the 4/5ths rule, (d) the number of repetitions that were resulted in violation of Ln_{adj}, (e) the number of repetitions that were resulted in violation of the Z_D test, and (f) the number of repetitions that were resulted in violation of the Z_{IR} test.

To compute SR_{diff}, I subtracted SR_{min} from SR_{maj}. Then, I summed all the results from each repetition and divided this sum by 1000 (total number of repetitions) to get the average difference between SR_{min} and SR_{maj}. To compute $IR_{average}$, first, I computed IR for each repetition and summed them. Then, I divided the sum, again, by 1000.

In order to compute the number of repetitions that were resulted in violation of the 4/5ths rule, first, I divided SR_{min} by SR_{maj} to calculate IR and compared it to the benchmark of .80. Values below .80 indicated violations of the 4/5ths rule. Secondly, I summed the number of repetitions indicating an IR below .80. The same procedure and benchmark were used to compute the number of repetitions that were resulted in violation of Ln_{adj}. First, I took the natural logarithm of the SR_{maj} [$\ln(SR_{maj})$] and the natural logarithm of the SR_{min} [$\ln(SR_{min})$]. Then, I computed IR_{Ln} by dividing $\ln(SR_{maj})$ by $\ln(SR_{min})$ and multiplied it by $\ln(0.8*SR_T + 2.63)$ to adjust to the changes in SR_T. The

next step was to compare the obtained value to the benchmark of .80 to find if AI was evident. As a last step, I summed, again, the number of repetitions resulted in a value lower than .80.

To find the number of repetitions that indicates a significant difference between SR_{maj} and SR_{min} for the Z_D test, first, I subtracted SR_{maj} from SR_{min} (SR_{min}-SR_{maj}). Then, I divided the obtained difference by its own standard error. The next step was to check if the obtained value was in the critical region for a two-tailed *z*-test with an alpha level of .05. The critical values for a two-tailed *z*-test with an alpha level of .05 are +1.96 and -1.96. If the obtained value was either lower than -1.96 or higher than 1.96, it indicated AI. As a final step, I summed the number of repetitions resulted in a value either lower than -1.96 or higher than 1.96. A similar procedure was used to calculate the number of repetitions indicating an IR significantly different from 1.0 for the Z_{IR} test. First, I took the natural logarithm of IR [$\ln(SR_{min}/SR_{maj})$], then, divided it by its own standard error. Following this, I checked, again, whether the resulted value was in the critical region, either lower than -1.96 or higher than 1.96. At the end, I summed the number of repetitions resulted in a value that fell in the critical region.

After obtaining the numbers of repetitions that resulted with a value indicating AI for each of these four measurements, I divided this number by 1000 (the total number of repetitions) to obtain the probability of observing AI for each measurement for a particular combination of SR_T, AR_{min}, and *n*.

When considering AI, it is intuitively assumed that the selection procedure used has an adverse impact on some pre-identified group (or minority group). Therefore, using a directional test seems to be more appropriate in many cases than using a bi-directional

test. However, not only the courts (*Casteneda v. Partida*, 1977) but also the federal regulations (OFCCP, 1993) have recommended the use of bi-directional (two-tailed) significance testing. To be in accord with these recommendations and the current use of these tests in simulation research (Morris & Lobsenz, 2000), I also used bi-directional test in this study.

2.2 Simulating Data with a Subgroup Mean Difference

The same procedure was repeated to calculate the probabilities of observing AI for a particular combination of AR_{min}, SR_T, and *n* when there is a subgroup mean difference ($d \neq .00$). SR_{diff} and $IR_{average}$ were computed again along with the probabilities of observing AI for each of the four AI measurements. The only difference was that minority applicant pools were generated by selecting minority applicants randomly from a population with a mean and a SD which are different from those of the majority population. In short, the parameters used to create minority population were different from the parameters used to create majority population. Table 2.1 presents the parameters used to generate minority populations for a particular *d* value. The majority population mean and SD were always set to be 100 and 15, respectively.

Table 2.1

Parameters Used to Generate Minority Populations for a Particular d Value

		d values			
Minority ratio	Parameters	.25	.50	.75	1.00
.12	μ_{min}	96.26	92.56	88.88	85.23
	σ_{min}	14.5	14	13.5	13
.20	μ_{min}	96.27	92.60	88.97	85.38
	σ_{min}	14.5	14	13.5	13
.30	μ_{min}	96.28	92.65	89.07	85.57
	σ_{min}	14.5	14	13.5	13

Note. Population means and standard deviations for the majority group were always equals to 100 and 15, respectively, through the simulations. When $d = .00$, the population mean and standard deviation for the minority group were equal to 100 and 15, respectively.

2.2.1 Factors Varied

There were three factors set to be varied in this study when generating the applicant pool. These factors include (a) the subgroup mean difference between majority and minority population from which the applicants were selected randomly, (b) the proportion of minority and majority group members in the applicant pool, and (c) applicant pool size (*n*).

The subgroup mean difference varied among .00, .25, .50, .75, and 1.00. The main reason choosing these values was to observe the behaviors of these four measurements when there was small ($d = .25$), medium ($d = .50$), large ($d = .75$ or $d = 1.00$), or no subgroup mean difference. Besides, the study by Hough et al. (2001), summarizing

standardized mean group (gender, ethnic, and age) differences across studies for cognitive ability, personality, and other predictor domains at both broadly and more narrowly defined construct levels, found that the standardized mean group difference on cognitive ability and personality ranged from none (d = .00) to 1.00 . For example, their result revealed that the standardized mean group difference between Blacks and Whites is 1.00 on general intelligence, .70 on quantitative ability, .60 on verbal ability, .50 on memory, .30 on mental processing speed, .21 on openness to experience, and .10 on extroversion. By setting d values at .00, .25, .50, .75, and 1.00, I tried to represent the majority of these observed subgroup mean difference at the population level.

It is evident that there were subgroup mean differences that were larger than 1.00 on some tests. For example, the subgroup mean difference between men and woman was 1.86 on muscular tension and 2.10 on muscular power (Hough et al., 2001). On a separate analysis, the simulation results indicated that the changes in the behaviors of these four measurements were negligible when the subgroup difference increased from 1.00 to 1.25 or to a larger value. Therefore, I concluded that varying the subgroup mean difference between .00 and 1.00 will be adequate to understand how these four measurements behave when assessing AI. Besides, there were some predictor domains where minority groups outperform majority group. For example, the subgroup mean difference between East Asians and Whites is .20 on cognitive ability (in favor of East Asians) and the subgroup mean difference between women and men is .64 on flexibility (in favor of women; Hough et al., 2001). Therefore, I would have generated minority population with parameters higher than those of majority population, created a subgroup mean difference in favor of minorities, and tested how these measurements behave when there was AI

against majority applicants. Although this might be a worthwhile endeavor on the grounds that *Uniform Guidelines* does not specify any particular group in defining AI, previous simulation research (Collins & Morris, 2008; Roth et al., 2006; Sackett & Roth, 1996) have only focused on the cases where there were AI against minority groups. I took the same perspective to be in line with these research and simulated cases where there was AI only against minority groups.

The second factor varied was the minority ratio in the simulated applicant pool. Three values of minority proportion were used. In the first condition minority proportion was set to .12 and in the second and third conditions it was set to .20 and .30, respectively. The minority proportions of .12 and .20 were also used by Roth et al. (2006) in their simulation. There were two reasons why I used these same values: (a) to make a comparison between the results of this study and the results of Roth et al.'s (2006) study in conditions where making such comparison was possible, and (b) to use values that did not deviate from the observed proportions in the population level. Regarding employment status of the civilian non-institutional population, the proportions of Blacks and Hispanics or Latinos in the workforce are about .10 and .13, respectively (U.S. Department of Labor & U.S. Bureau of Labor Statistics, 2011). The minority proportion in the first condition here reflects these statistics. In line with Roth et al.'s (2006) reasoning, in the second condition, I used .20 as the proportion of minority applicants to allow generalizations to other selection settings that might be observed in the applied world. Although the minority proportion of .12 reflects the Black and Hispanic U.S. workforce in general, there are some geographical areas where the proportion of either of these two subgroups can be larger than .12 or .20. For example, Hispanics constitutes 38% of the population in

Texas[17]. Thus, observing minority applicant ratios around .30 is within the realm of possibility. To address this possibility, I also compared the behavior of AI measurements in conditions where the minority applicant ratio is set to .30. Results of these simulations are presented in Figure 3.1.3 through Figure 3.5.3 in Appendix B.

The third factor varied was the applicant pool size. Morris and Lobsenz (2000) set their applicant number to vary between 100 and 1000 when they compare the statistical power of the Z_{IR} test and the Z_D tests and Roth et al. (2006) set their applicant number to be 200, 400, and 2000 in their simulations. To be in an accord with these studies, I set the applicant pool size to 200, 400, 1000, 2000, and 5000. Although neither of these two simulations set their applicant pool size as large as 5000, I included this value to evaluate the behaviors of AI measurements when the applicant pool size is very large. Studying the relationship between recruitment source, applicant quality, and hire performance, Kirnan, Farley, and Geisinger (1989) examined the applicant data from a major insurance company in the year 1981. The number of applicants for the insurance agent position was reported as 20,576 in this study. Considering this example in line with the increase in the number of large public and private organizations and the large difference between workforce demand and supply in times of tight economies, having an applicant pool size of 5000 was not an unrealistic assumption in AI analysis. Therefore, I decided to include conditions where the applicant pool size was as large as 5000.

[17] http://quickfacts.census.gov/qfd/states/48000.html

Chapter 3

Results

One of the objectives of this study was to examine how each of the four AI measurements (the 4/5ths rule, Ln_{adj}, the Z_{IR} test, and the Z_D test) behave when there was no subgroup mean difference (d = .00) on the selection test used. Figure 3.1.1 and Figure 3.1.2 provide (a) the probabilities of obtaining AI for each of the four AI measurements, (b) average difference between minority and majority selection ratios, and (c) average sample impact ratio, respectively for AR_{min} of .12 and AR_{min} of .20, in cases where d was set to .00. No subgroup mean difference denotes no AI. Therefore, the probabilities presented in Figure 3.1.1 and Figure 3.1.2 are to be considered as Type I error rates.

Figure 3.1.1 illustrates that the probabilities of observing AI are higher for the 4/5ths rule and Ln_{adj} than for the Z_{IR} test and the Z_D test when the applicant pool size (n) is small (either 200 or 400). When n = 200, the 4/5ths rule and Ln_{adj} indicate AI, on the average, 14 and 22 percent of the time, respectively. The same percentages are 7 and 4, respectively for the Z_{IR} test and the Z_D test. These percentages indicate that Type I error rates for the 4/5ths rule and Ln_{adj} are well-above the acceptable alpha level of .05 while Type I error rates for the tests of significance are within an acceptable range when n = 200. When n = 400, the 4/5ths rule indicates AI, on the average, 9 percent of the time. This was still somewhat higher than the acceptable alpha level. The average percentages for the other three AI measurements are relatively small: 7 for Ln_{adj} and 6 for both the Z_{IR} test and the Z_D test.

The other noticeable finding is that Type I error rates for the tests of significance are relatively equal across conditions while these error rates for the 4/5ths rule and Ln_{adj} are noticeably varied for different conditions. For example when n = 200, Type I error rates for the 4/5ths rule are .29, .14, and .01, respectively for the selection ratios of .10, .50, and .90. The same error rates for Ln_{adj} are .29, .14, and .28, respectively, again for the selection ratios of .10, .50, and .90. When n increases to 400, Type I error rates decreases both for the 4/5ths rule and Ln_{adj}. But, behaving less conservatively when SR_T is small, Type I error rates for the 4/5ths rule is .27 for the selection ratio of .10, well-above the observed Type I error rate (which is .14) for Ln_{adj} in the same condition. Type I error rates are not a real concern for the tests of significance when applicant pool size is small, because the error rates for these tests are ranged between .04 and .08 across different total selection ratios.

As the applicant pool size increases, however, the Type I error rates for the 4/5ths rule and the Ln_{adj} rule decrease sharply. When n is equal to or higher than 1000, Type I error rates for Ln_{adj} are, on the average, within the acceptable range. These error rates are almost zero when n is equal to or higher than 2000. Type I error rates for the 4/5ths rule are also within the acceptable range when SR_T is equal to or higher than .30. Even with n as large as 1000 and 2000, the 4/5ths rule indicates Type I error about 20% and 13% of the time, when SR_T = .10. An increase in the applicant pool size does not affect the amount of Type I error rates observed for the tests of significance. Although, the Z_D test produces, on the average, less Type I error rates than the Z_{IR} test, the Type I error rates for these tests of significance are comparable and within the acceptable range.

Figure 3.1.1 Simulation Results Where d = .00 (no AI) and AR_{min} = .12

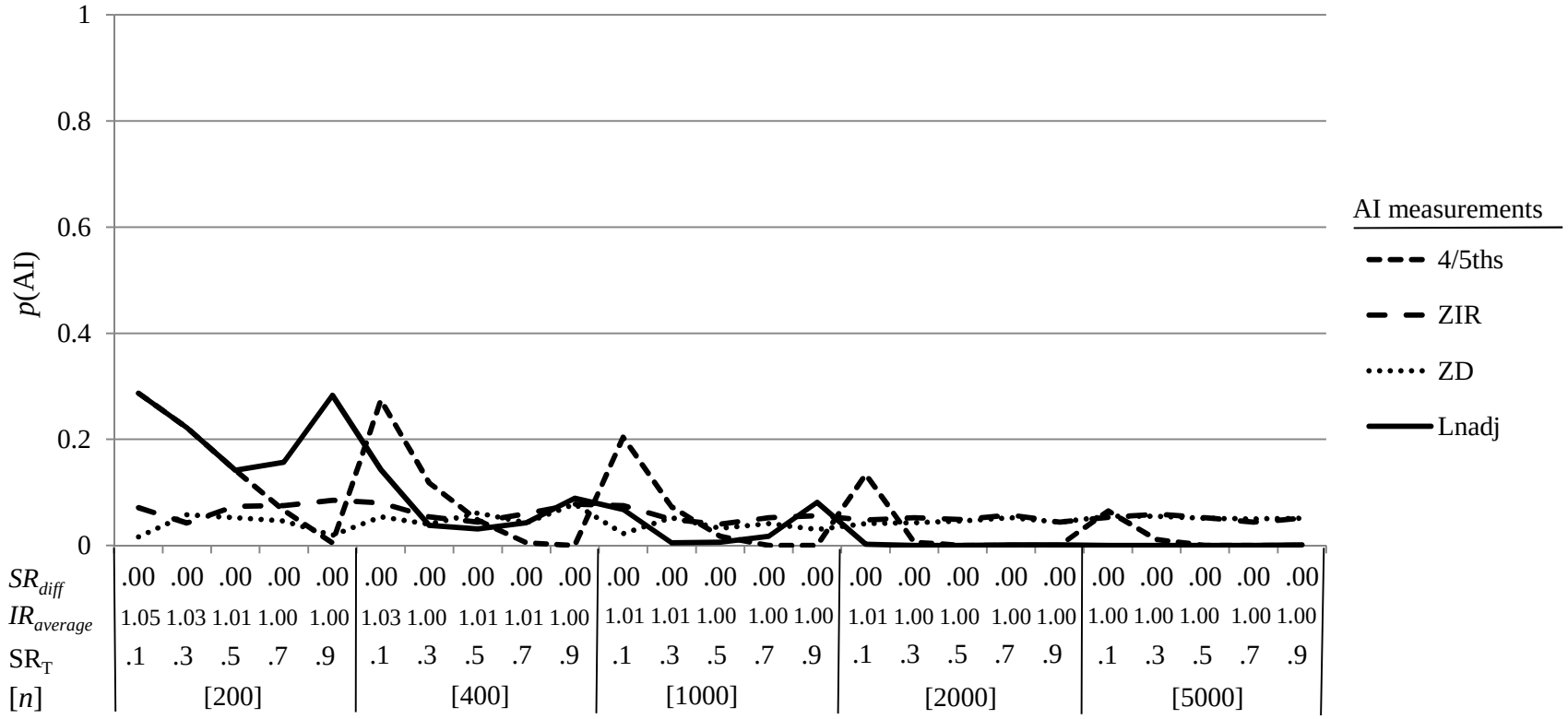


Figure 3.1.1 Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (SR_{diff}), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [n] when d = .00 and minority applicant ratio (AR_{min})= .12.

Figure 3.1.2 provides Type I error rates when the minority ratio is set to .20. Increasing AR_{min} from .12 to .20 does not profoundly change the Type I error rates for the discussed AI measurements. When SR_T and applicant pool size are small, Type I error rates for the 4/5ths rule and Ln_{adj} are still above the acceptable alpha level of .05. For example, with a SR_T of .10, the 4/5ths rule indicates Type I error 40% and 26% of the time, respectively for the applicant pool sizes of 200 and 400. These percentages are decreased to 17% and to 15% for the same conditions when Ln_{adj} is used to assess AI. Increasing SR_T to .90, Type I error rates constituted no problem for the 4/5ths rule while they became a real concern for Ln_{adj}. When SR_T is set to .90, Type I error rates for Ln_{adj} are .19 and .15, respectively for the applicant pool sizes of 200 and 400.

Increasing applicant pool size, again, leads to a decrease in Type I error rates for Ln_{adj} and the 4/5ths rule. Regardless of the magnitude of SR_T, Type I error rates are within the acceptable range for Ln_{adj} when *n* is large ($n \geq 1000$). For the 4/5ths rule, Type I error rates were also within the acceptable range except the conditions where SR_T is set to .10. Even with a large applicant sample, the 4/5ths rule still produces noticeable Type I error rates when SR_T is equal to .10. These error rates are .17 and .12, respectively for the applicant pool size of 1000 and 2000.

The tests of significance again produce acceptable Type I error rates across selection ratios regardless of the applicant pool size when AR_{min} is set to .20. In the meantime, results again indicate that Type I error rates for the Z_D test are, on the average, lower than Type I error rates for the Z_{IR} test.

Figure 3.1.2 Simulation Results Where d = .00 (no AI) and AR_{min} = .20

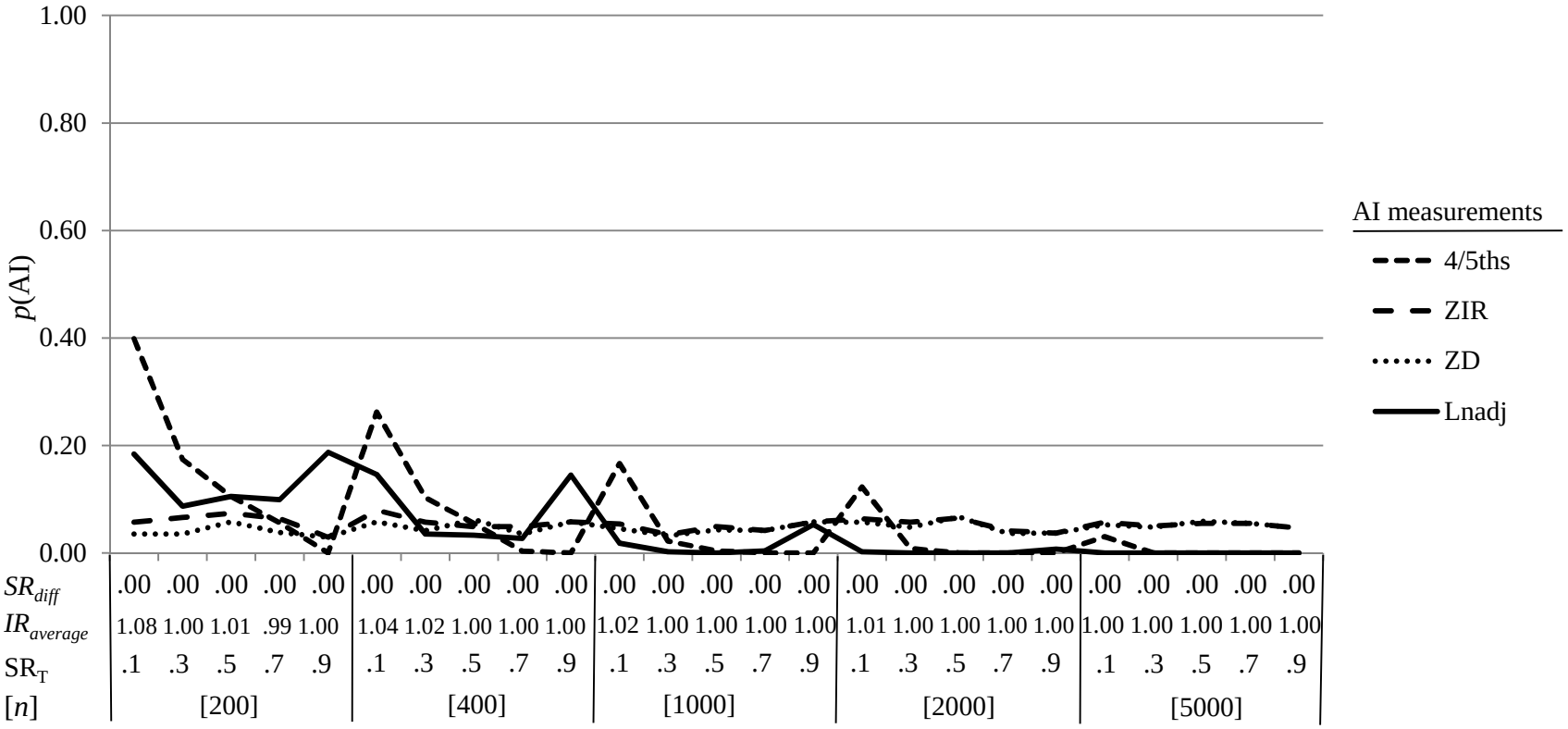


Figure 3.1.2 Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (SR_{diff}), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [n] when d = .00 and minority applicant ratio (AR_{min})= .20.

Comparing the behaviors of these four AI measurements in cases where $d = .00$, the tests of significance usually produce lower Type I error rates than the 4/5ths rule and Ln_{adj}. More importantly, the distributions of error rates across conditions are more symmetric for the Z_{IR} test and the Z_D tests than those of the 4/5ths rule and Ln_{adj}. Considering the two tests of practical significance (the 4/5ths rule and Ln_{adj}), two conclusions can be reached by examining Figures 6.1 and Figure 3.1.2. First, Type I error rates for Ln_{adj} are, on the average, lower than those of the 4/5ths rule. Second, the distribution of error rates is more symmetrical for Ln_{adj} than it is for the 4/5ths rule. In short, although Ln_{adj} performs poorer than both of the tests of significance in no AI conditions when applicant pool size is small, it outperforms the tests of significance and the 4/5ths rule when the applicant pool size is large ($n \geq 1000$).

In addition to varying factors (SR_T, AR_{min}, d, and n) related to personnel selection outcome and evaluating the behaviors of AI measurements regarding how they responded to these variations, I computed average differences between SR_{min} and SR_{maj} (SR_{diff}) and average sample impact ratio ($IR_{average}$) and used these variables as benchmarks to compare the behaviors of AI measurements. It was expected that in cases where a large SR_{diff} in favor of SR_{maj} or a small $IR_{average}$ was observed, AI measurements would indicate AI. Examining Figure 3.1.1 and Figure 3.1.2, selection ratio differences are equal to zero while average sample impact ratios were around 1.00. Because neither selection ratio differences are large nor IR averages are small; AI measurements that are, on the average, less likely to indicate AI outperforms other AI measurements in these scenarios.

Although selection ratio differences presented in Figure 3.1.1 and Figure 3.1.2 are always around .00, there are some changes in average impact ratios as SR_T is varied. The

most noticeable changes (as a deviation from 1.00) in average sample impact ratios are observed in conditions where both applicant pool size and SR_T are small. With an applicant pool size of 200 and SR_T of .10, the observed $IR_{average}$ are 1.05 and 1.08, respectively in Figure 3.1.1 and Figure 3.1.2. Obtaining an $IR_{average}$ higher than 1.00 means that SR_{min} is, on the average, higher than SR_{maj} in these conditions. However, the results of simulations reveal that the 4/5ths rule and Ln_{adj} are violated more frequently in these particular conditions. This peculiarity in the findings indicates the tremendous effect of hiring or not hiring one more minority applicant when both sample size and selection ratios are small. For example, when SR_T = .10, AR_{min} = .12, and n = 200, hiring 2 minority applicant would result in an IR of .81, hiring one more minority applicant would result in an IR of 1.22. Being aware of this effect, *Uniform Guidelines* instituted "*N* of 1" or "flip-flop" rule (EEOC et al., 1978). This rule basically allows one to assume that the organization hires one more minority group member and one less majority group member. Then, based on this assumption, if the order of the selection ratios is reversed ($SR_{min} > SR_{maj}$), it is understood that AI didn't occur. In short, these results also support the use of "flip-flop" rule in cases where hiring one more minority applicant changes the order of the selection ratios.

The interpretation of simulation results is more complicated when there is a small *d* value. It is reasonable to expect an increase in the probabilities of observing AI for all of the four AI measurements when *d* value increases from .00 to .25. To what extent is this increase reasonable, however, becomes a question of interest. Although providing an exact answer to this question is outside the scope of this study, I can easily argue that when the probabilities of observing AI approach 1.00, it is reasonable to expect an

increase in Type I error rates. Knowing that practical significance is the main concern in the *Uniform Guidelines*, a small *d* value (.25) producing large effects (e.g., causing AI at 90% of the time) does not comply with this concern. Thus, the increase in probabilities of observing AI for these AI measurements should be small. Nevertheless, I can compare the behaviors of the four AI measurements when $d = .25$ and discuss the noticeable trends in their behaviors across all the combinations of SR_T and *n*.

The study by Sackett and Ellingson (1997) revealed that small *d* values could cause violation of the 4/5ths rule at selection ratios that were frequently observed. As Figure 3.2.1 presents, the simulation results indicate that when $d = .25$, the 4/5ths rule is violated more than 80% of the time for small SR_T in conditions where applicant pool size is large ($n \geq 1000$) and AR_{min} is .12. For example, with a SR_T of .10, the probabilities of observing AI for the 4/5ths rule are .90 and .98, respectively for the applicant pool sizes of 2000 and 5000. Conversely, when SR_T approaches 1.00, the probabilities of observing AI for the 4/5ths rule decrease to as low as .00. Holding selection ratio constant at .90, the 4/5ths rule indicates no AI across applicant pool sizes except the condition where *n* is set to 200.

When the applicant pool size is small, the 4/5ths rule is, on the average, less likely to indicate AI. Although the 4/5ths rule indicates AI about 98% of the time in condition where n = 5000 and $SR_T = .10$, it only indicates AI about 70% and 56% of the time when *n* decreases to 400 and 200, respectively. These noticeable changes in the probabilities of observing AI, as graphed in the Figure 3.2.1, illustrate the sensitivity of the 4/5ths rule to changes not only in SR_T but also in applicant pool size.

The results for Ln_{adj} indicate that this measurement, compared to the 4/5ths rule, is less sensitive to the changes in SR_T and applicant pool size. Although the probabilities for the 4/5ths rule range from .00 to .97, these probabilities range from .25 to .56 for Ln_{adj} across all the conditions where d = .25 and AR_{min} = .12. Higher rates for the violation of Ln_{adj} are observed, again, when applicant pool size is small and SR_T is either too small or too large. For example, holding applicant pool size constant at 200, Ln_{adj} is violated 56%, 41%, and 45% of the time, respectively for the SR_Ts of .10, .50, and .90. An increase in applicant pool size leads to a decrease in the probabilities of observing AI for Ln_{adj}. But, the effect of changes in applicant pool size for Ln_{adj} is relatively small.

Figure 3.2.1. Simulation Results Where d = .25 and AR_{min} = .12

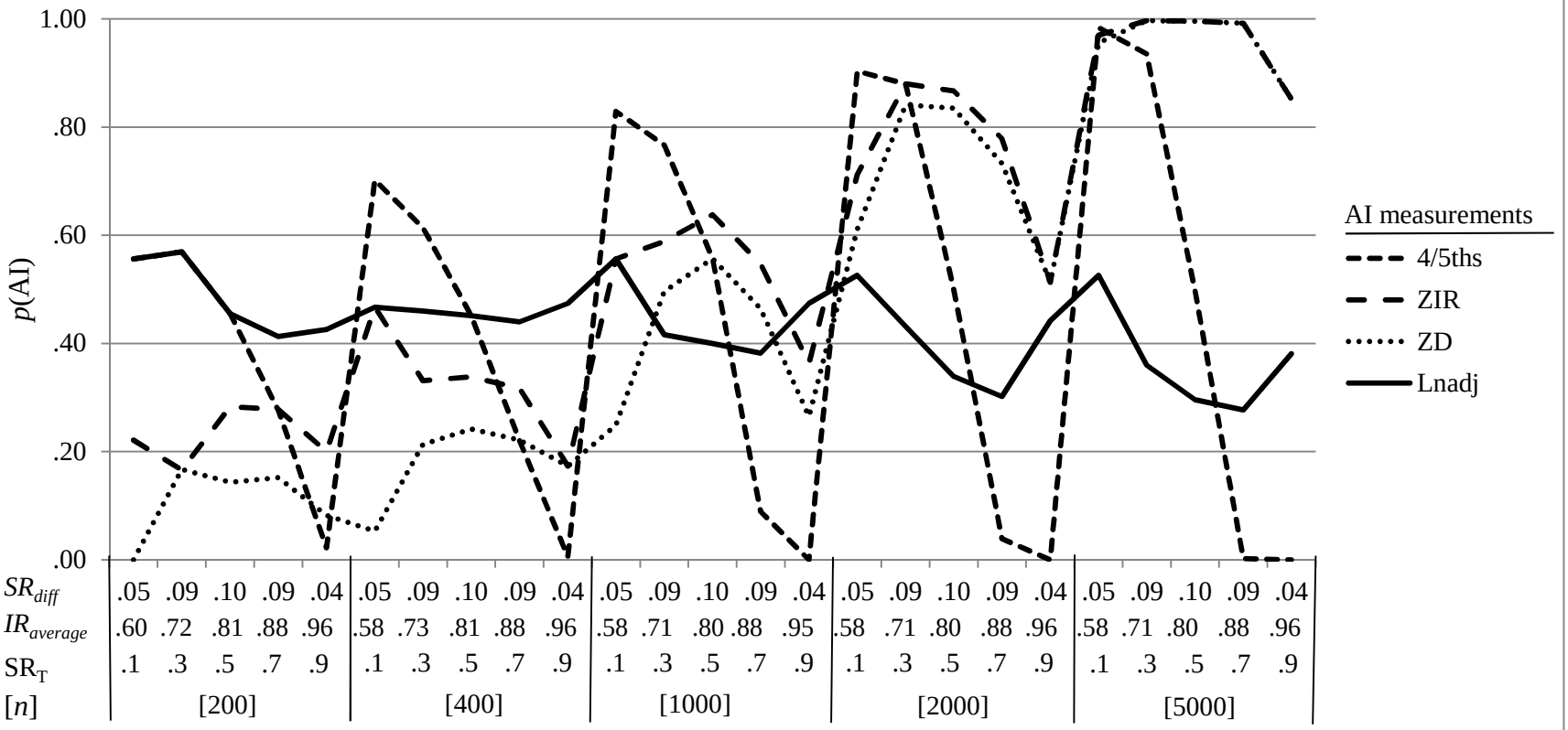


Figure 3.2.1. Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (SR_{diff}), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [n] when d = .25 and minority applicant ratio (AR_{min})= .12.

The results for the Z_D test and the Z_{IR} tests are to be discussed in line with the issue of statistical power. It is a well-known issue that as the sample size increases the power of statistical tests increases, and this increase in power leads to higher likelihoods of obtaining significant results. This issue is clearly demonstrated in Figure 3.2.1 and Figure 3.2.2. When the applicant pool sizes are 200 and 400, the probabilities of observing AI are lower than .40 for these tests of significance. On the other hand, these probabilities became as large as 1.00 as the applicant pool size increases to 2000 and 5000. The results also reveal that significance tests are sensitive to changes in SR_T. For example, holding applicant pool size constant at 2000, the Z_{IR} test indicates AI 86% and 50% of the time, respectively for the SR_T of .50 and .90. The Z_D test indicates AI 83% and 50% of the time, respectively for those same conditions.

The results of the simulations where d = .25 and AR_{min} = .20 are presented in Figure 3.2.2. The behavioral patterns of these four AI measurements do not vary noticeably after minority ratio increase from .12 to .20. However, on the average, a .06 increase is detected in the probabilities of observing AI for the Z_{IR} test and a .10 increase is detected for the Z_D test. The probabilities for the 4/5ths rule and Ln_{adj} stayed relatively stable.

Figure 3.2.1 and Figure 3.2.2 illustrate that changes not in sample size but in SR_T have a direct effect on SR_{diff} and $IR_{average}$. The relationship between SR_T and $IR_{average}$ is linear; $IR_{average}$ increases as SR_T increases. However, the relationship between SR_T and SR_{diff} is non-linear (inverted-U); SR_{diff} increases as SR_T approaches .50 and decreases as it approaches .00 and 1.00. The same trend is observed across all applicant sample sizes.

Considering SR_{diff} and $IR_{average,}$ the behavior of the 4/5ths rule is best predicted by $IR_{average}$; holding applicant pool size constant, the violation of the 4/5ths rule decreases as $IR_{average}$ increases. A similar trend is observed for Ln_{adj} in the conditions where $SR_T \leq .70$. When $SR_T = .90$, an increase in $IR_{average}$ is accompanied by an increment in the violation of Ln_{adj}. The behavior of Ln_{adj} is also fairly predicted by SR_{diff}; similar SR_{diff} indicates similar rates of violation for Ln_{adj} in conditions where SR_T is neither small (.10) nor large (.90). A relatively small SR_{diff} is more likely to indicate the violation of Ln_{adj} when SR_T is either small or large.

Although observing an increase in the violation of Ln_{adj} along with a decrease in SR_{diff} in conditions where SR_T is either .10 or .90 seems contradictory, the ceiling effect in SR_{diff} observed in these conditions should be taken into consideration for a comprehensive understanding. For example, setting *n* to 1000 and AR_{min} to .20, the maximum value SR_{diff} can get is .13 when SR_T equals to .10. However, the possible maximum SR_{diff} values are .38, .63, and .88 respectively for the selection ratios of .30, .50, and .70; as SR_T increases, the maximum value for SR_{diff} increases. This positive linear relationship between SR_T and SR_{diff} becomes negative in conditions where SR_T exceeds AR_{maj}. Thus, the possible maximum value for SR_{diff} decreases to .50 when SR_T increases to .90. Note that AR_{maj} (.80) is smaller than SR_T (.90) in this last condition. Because of this ceiling effect a small SR_{diff} produces relatively large affects (an increase in the likelihood of violation of Ln_{adj}) when SR_T is either small or large.

The behaviors of the tests of significance are not predictable by $IR_{average}$, but by SR_{diff}. Holding *n* constant, an increase in SR_{diff} indicates an increase in the probabilities of obtaining a significant results (finding evidence for AI). When the variations in *n* are

taken into account, the relationship between the behaviors of the tests of significance and SR_{diff} gets weaker. Referring to Figure 3.2.2, the same SR_{diff} of .05 indicates AI 27% and 7% of the time, respectively for the Z_{IR} test and the Z_D test when n = 200. Increasing n to 5000, the same SR_{diff} indicates AI 100% of the time for both the Z_{IR} test and the Z_D test. Holding SR_{diff} constant, the increase in percentage of observing AI as n gets larger is associated with the statistical power. As n gets larger and larger, the same effect size (SR_{diff}) is more likely to be found as significant.

Figure 3.2.2 Simulation Results Where d = .25 and AR_{min} = .20

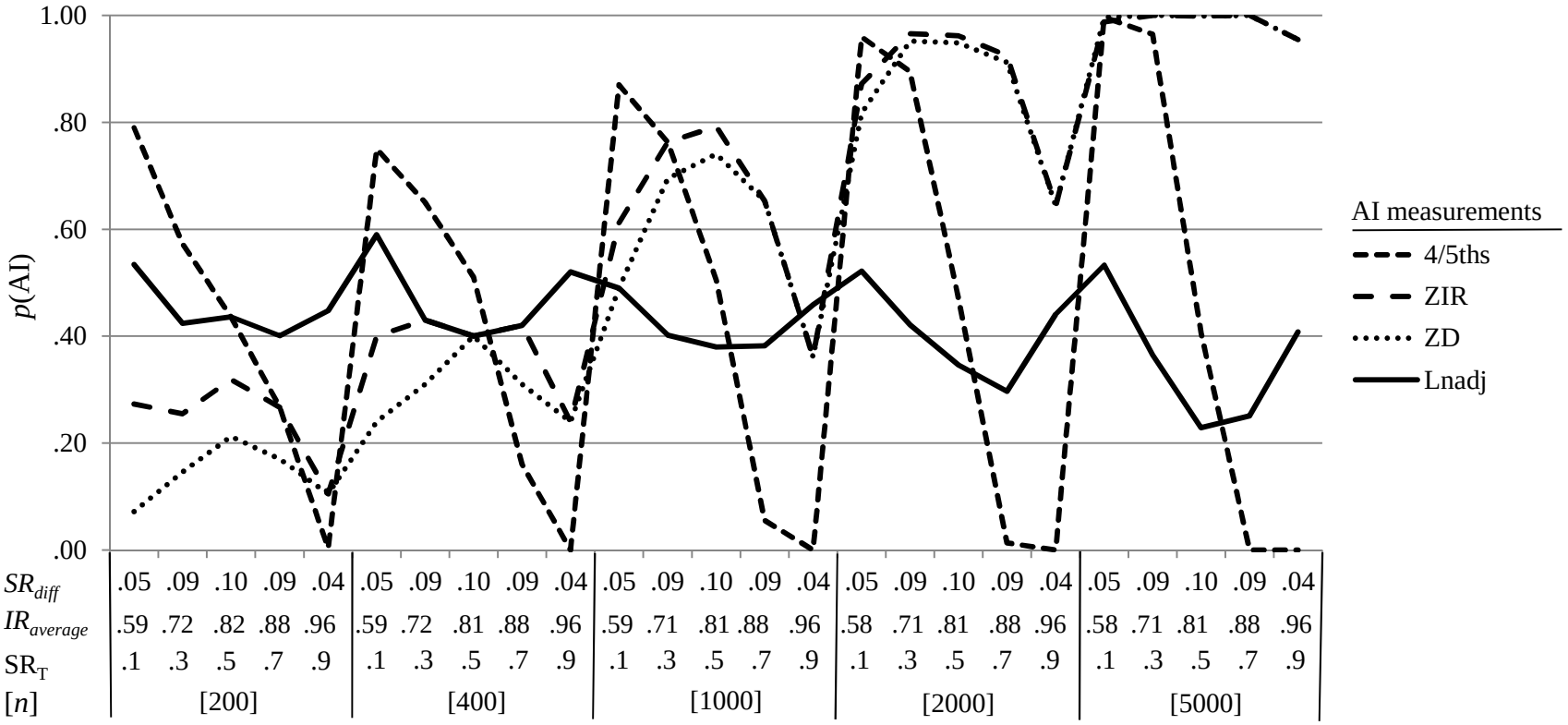


Figure 3.2.2 Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (SR_{diff}), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [n] when d = .25 and minority applicant ratio (AR_{min})= .20.

Examining both Figure 3.2.1 and Figure 3.2.2, the results indicate that the pattern observed in the behavior of Ln_{adj} is distinct from the pattern observed in the behaviors of the other three measurements at least in two ways. First, the probabilities of observing AI for Ln_{adj} stay relatively stable across all the combinations of SR_T and *n*. The fluctuations in these probabilities are more profound for the 4/5ths rule and the tests of significance than they are for Ln_{adj}. This indicates that Ln_{adj} is less sensitive to the changes in SR_T and *n* than the other three measurements are. Second, Ln_{adj}, on the average, is less likely to indicate AI compared to the other three AI measurements. Being less likely to indicate AI in conditions where *d* is small, it can be argued that Ln_{adj} is in accord with the notion of practical significance as discussed in the *Uniform Guidelines*.

The results for the conditions where *d* is set to .50 are presented in Figure 3.3.1 and Figure 3.3..2 for the minority applicant ratios of .12 and .20, respectively. As the results suggested Ln_{adj} is, on the average, more likely to indicate AI compared to the other three AI measurements when there is a moderate *d* value. The behaviors of Ln_{adj} show similar trends both in Figure 3.3.1 and Figure 3.3.2. Regardless of the magnitude of SR_T, Ln_{adj} indicates AI about 100% of the time when *n* is equal to or higher than 1000. A similar conclusion is evident for the tests of significance. These probabilities for Ln_{adj} and the tests of significance decrease when *n* drops below 1000. But, the decrease for Ln_{adj} is relatively small in magnitude compared to decrease for the tests of significance. Holding *n* constant at 200, Ln_{adj} indicates AI, on the average, about 78% of the time; the percentages are 53% and 34%, respectively for the Z_{IR} test and the Z_D test for the same condition.

Figure 3.3.1 Simulation Results Where d = .50 and AR_{min} = .12

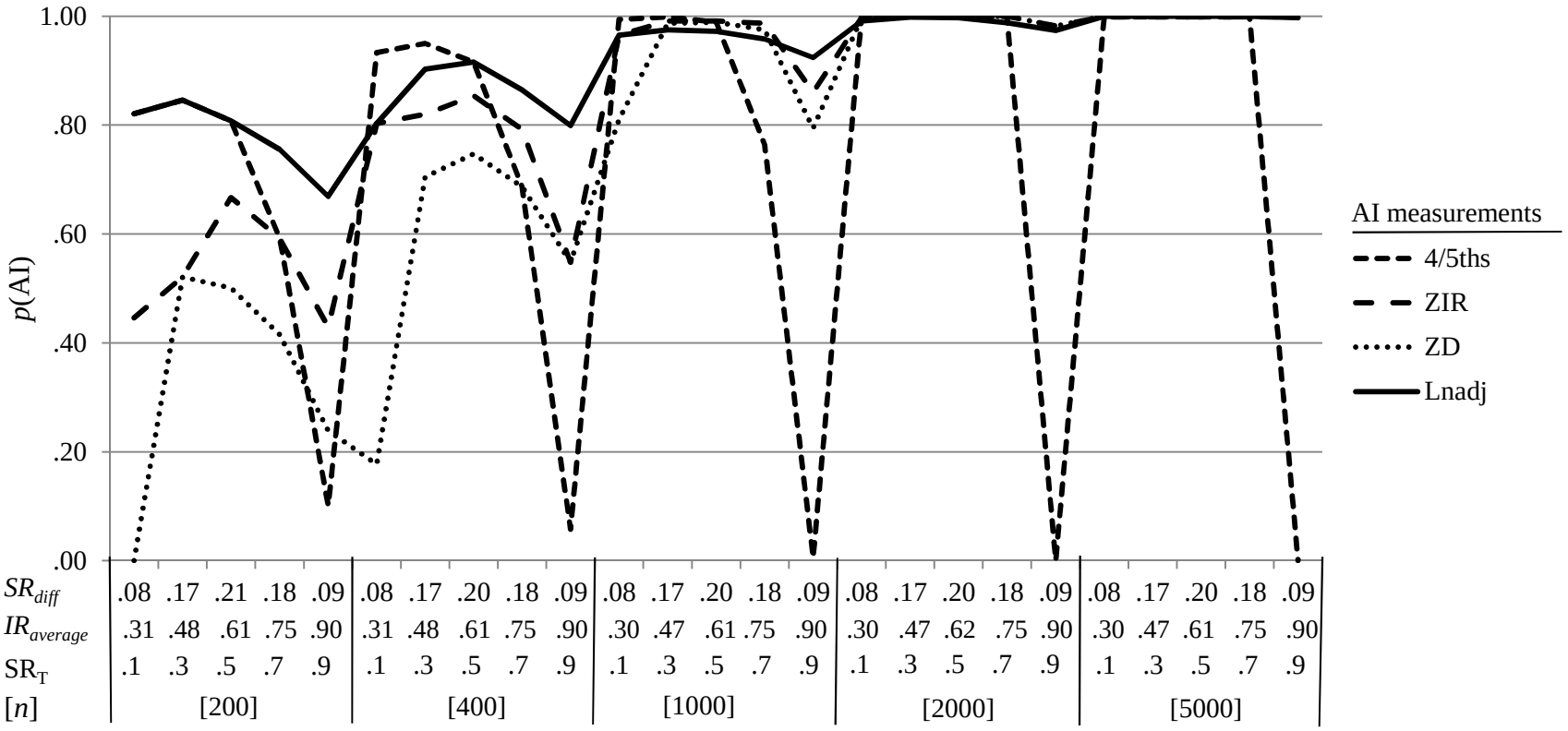


Figure 3.3.1 Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (SR_{diff}), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [n] when d = .50 and minority applicant ratio (AR_{min}) = .12.

Examining both Figure 3.3.1 and Figure 3.3.2, the sensitivity of the 4/5ths rule to the changes in SR_T is clearly observed. For example holding the applicant pool size constant at 200, a SR_T of .30 indicates AI 86% of the time whereas a SR_T of .90 indicated AI 11% of the time. The behavior of the 4/5ths rule, however, is not affected as much by the changes in *n* when *d* is moderate. The same trend of indicating higher rates of AI with small SR_T and lower rates of AI with higher SR_T is observable across different applicant pool sizes.

The results for the Z_{IR} test and the Z_D tests again reveal that these significance tests are not only sensitive to the changes in applicant pool size but also to the changes in SR_T. With a relatively small applicant pool size, the probabilities of observing AI change markedly as SR_T is set to vary from .10 to .90. For the Z_{IR} test, for example, the probability of observing AI increased from .55 to .85 as the SR_T decreased from .90 to .50 in the cases where the applicant pool size is set to 400 (refer to Figure 3.3.1). A similar trend is observed in the behavior of the Z_D test. The results also reveal that the Z_{IR} test, on the average, is more likely to indicate AI than the Z_D test. The results, in that sense, confirm Morris and Lobsenz (2000) who discussed that the Z_{IR} test had slightly better statistical power than the Z_D test under some conditions.

Figure 3.3.2 Simulation Results Where d = .50 and AR_{min} = .20

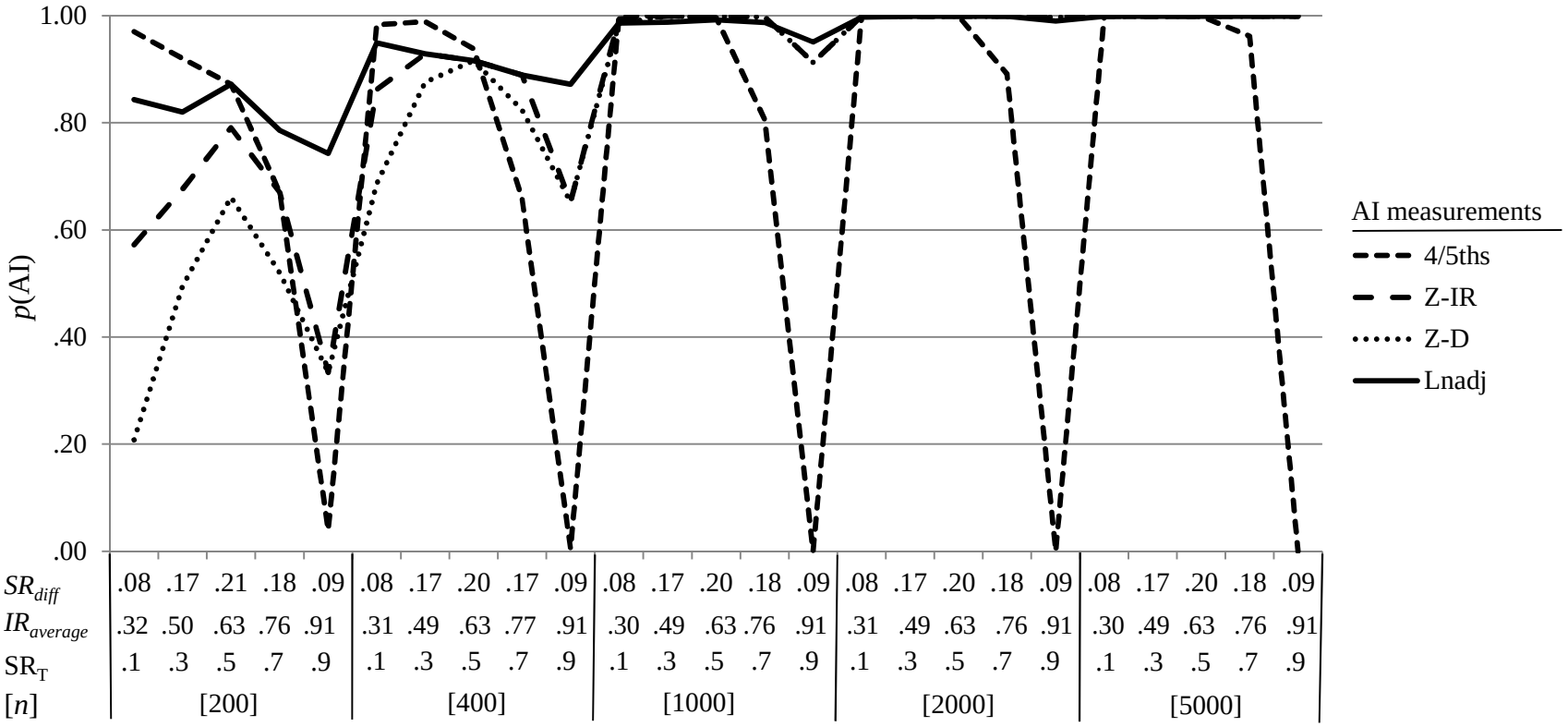


Figure 3.3.2 Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (SR_{diff}), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [n] when d = .50 and minority applicant ratio (AR_{min}) = .20.

The results presented in both Figure 3.3.1 and Figure 3.3.2 clearly show that Ln_{adj} produces relatively stable probabilities of observing AI across all combinations of SR_T and *n*. The results for the tests of significance are highly dependent on the changes in *n*. When *n* is small, the results for the tests of significance are less likely to indicate AI with a low or high SR_T and more likely to indicate AI with a moderate SR_T. Higher selection ratios continue to be a problem for the 4/5ths rule. Even though, there is a meaningful subgroup mean difference (*d* = .50), the probability of observing AI for the 4/5ths rule decreases to as low as .00 when SR_T approaches 1.00.

Examining SR_{diff} and $IR_{average}$ values, the behavior of the 4/5ths rule, again, is best predicted by $IR_{average}$ and the behaviors of the tests of significance are best predicted by SR_{diff}. The behavior of Ln_{adj} is best predicted by $IR_{average}$ when sample size is small and by SR_{diff} when sample size is large. The results also indicate that the tests of significance fail to indicate AI more than half of the time even when $IR_{average}$ as small as .31is present in conditions where both SR_T and *n* are small.

Figures 9.1 and Figure 9.2 present simulation results for the cases where *d* = .75. Having such a large *d* value, it is expected that these measurements should indicate AI almost in all cases. Although the results confirm this expectation, there are still some concerns regarding the 4/5ths rule and the Z_D test. For example, with an applicant pool size of 2000 and a minority applicant ratio of .12, the 4/5ths rule indicates AI only 13% of the time when SR_T is .90. Having slightly less statistical power than the Z_{IR} test, small applicant pool sizes still constitute a problem for the Z_D test. Setting SR_T to .10 and AR_{min} to .12, the Z_D test indicates AI 0% and 49% of the time, respectively for the applicant pool sizes of 200 and 400. The results are fairly stable for both Ln_{adj} and the Z_{IR} test.

However, Ln_{adj} outperforms the Z_{IR} test by indicating higher probabilities of observing AI when the applicant pool size is smaller than 1000. Notice that when there is a large subgroup mean difference, indicating higher probabilities of observing AI is a desirable outcome.

An increase in the minority applicant ratio from .12 to .20 does not cause a noticeable change in the behaviors of these AI measurements when d = .75. The same behavioral pattern is observed for the 4/5ths rule. As SR_T approaches 1.00, the probabilities of observing AI approach .00. Small applicant pool size is still a problem for both the Z_{IR} test and the Z_D test. The most visible change is observed in the behavior of the Z_D test in the case where applicant pool size is 200 and SR_T is .10. Although the Z_D test indicated no AI at all in this particular case when AR_{min} was .12, it indicated AI 50% of the time when AR_{min} increased to .20. Overall, behaving relatively stable even when the applicant pool size is as small as 200, Ln_{adj} outperforms the other AI measurements in cases where d = .75.

Figure 3.4.1 Simulation Results Where d = .75 and AR_{min} = .12

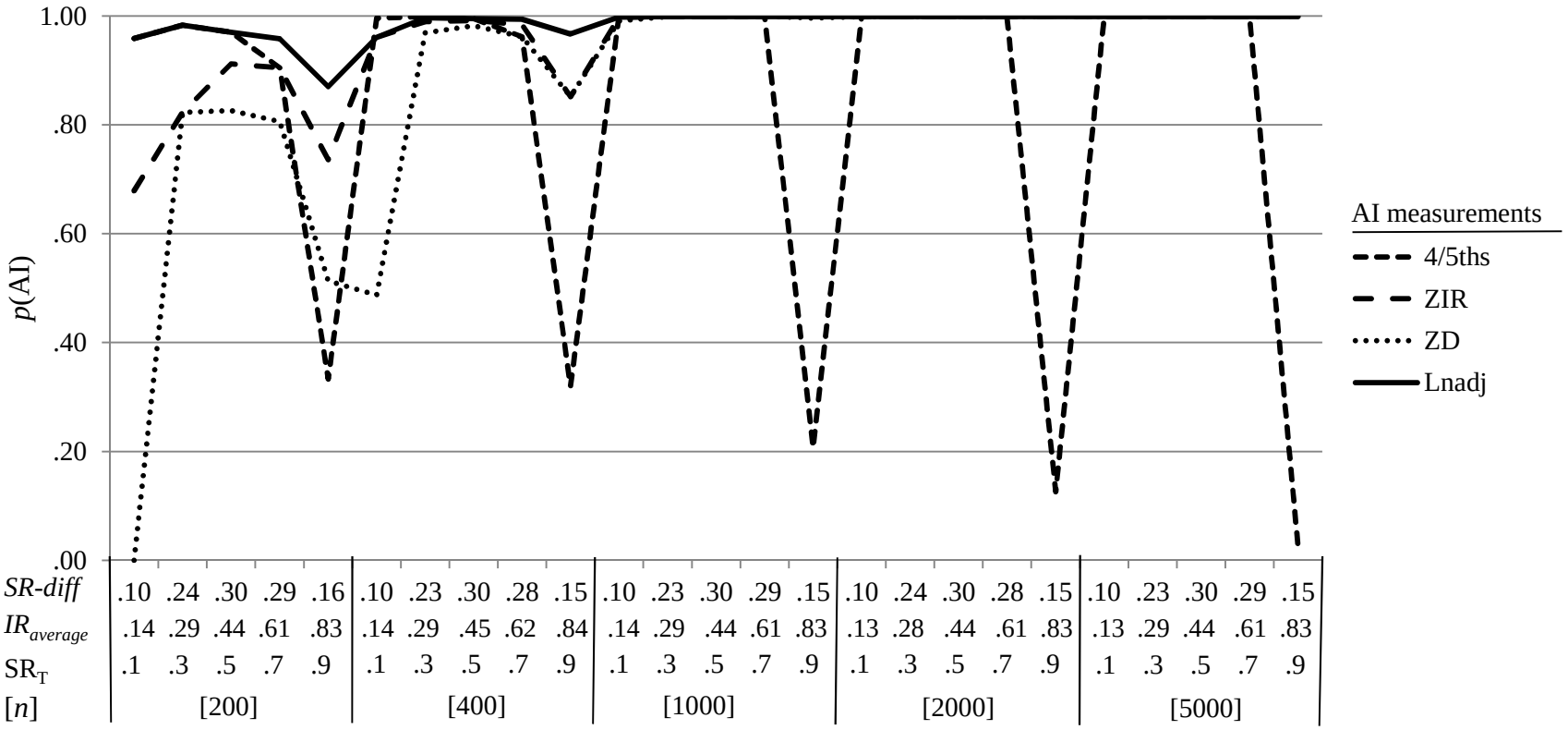


Figure 3.4.1 Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (*SR-diff*), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [*n*] when d = .75 and minority applicant ratio (AR_{min}) = .12.

Figure 3.4.2 Simulation Results Where d = .75 and AR_{min} = .20

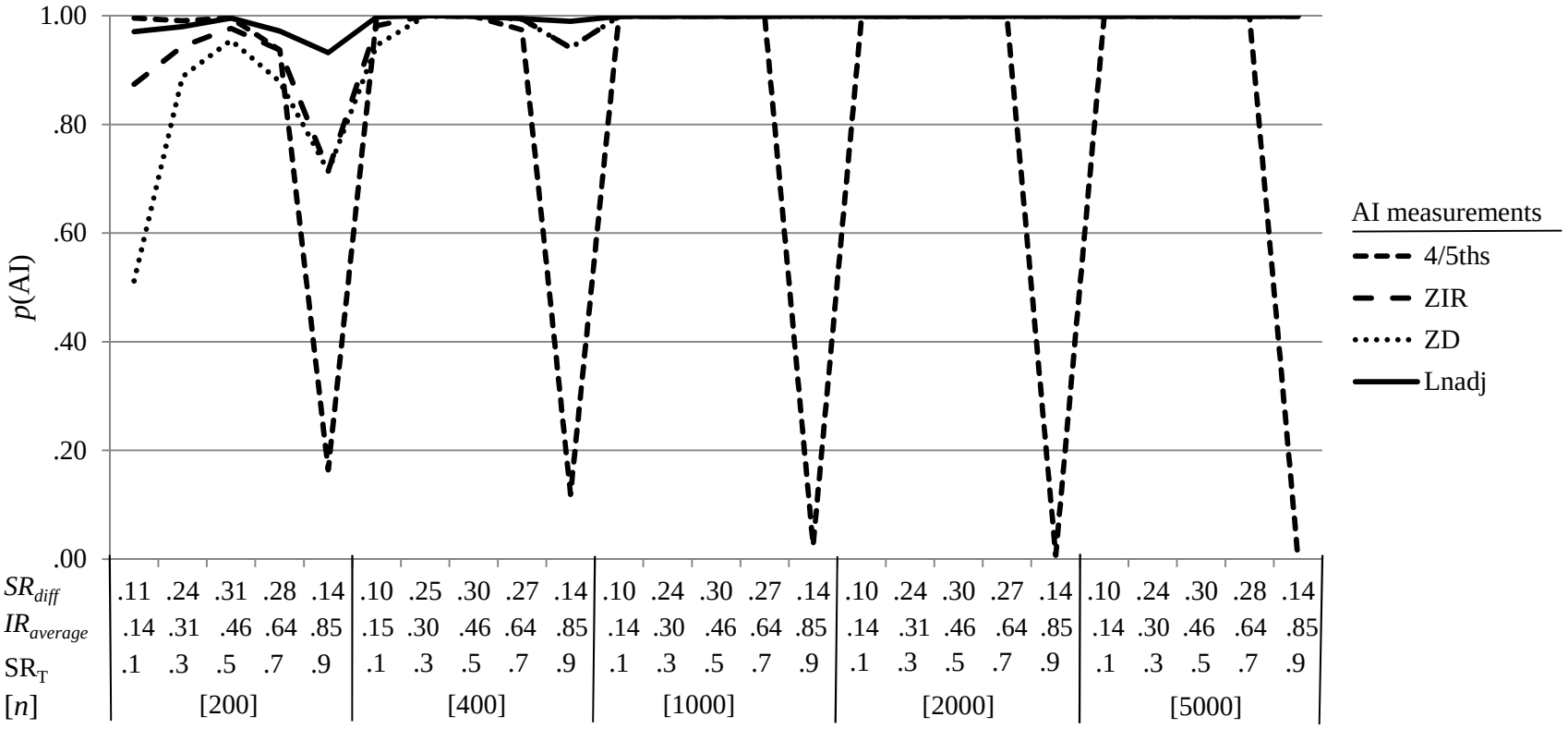


Figure 3.4.2 Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (SR_{diff}), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [n] when d = .75 and minority applicant ratio (AR_{min}) = .20.

Examining SR_{diff} and $IR_{average}$ in Figure 3.4.1 and Figure 3.4.2, there are two noticeable findings that worth to discuss. First, the 4/5ths rule fails to indicate AI most of the time when SR_{diff} as large as .15 is evident in cases where SR_T = .90. As it is discussed before, the 4/5ths rule behaves conservatively as SR_T approaches 1.00. Failing to indicate AI when there is a large SR_{diff}, the appropriateness of the 4/5ths rule as an AI measurement became questionable, especially in cases where applicant pool size is large. For example when n = 5000, AR_{min} = .12, and SR_T = .90, a SR_{diff} of .15 means that 80 more minorities out of 600 minority applicants and 80 less majorities out of 4400 majority applicants are need to be hired to obtain a SR_{diff} of .00. Nevertheless, the 4/5ths rule indicates that hiring 80 less minority applicant out of 600 is not practically significant: a conclusion that could be easily argued against. The OFCCP's concern for the 4/5ths rule's aptness to assess AI when a large number of hiring is made is underlined here. Remember that the OFCCP recommends using statistical and practical significance tests rather than the 4/5ths rule to assess AI when there is a large number of hiring (OFCCP, 1993). Second, the Z_D test fails to indicate AI when $IR_{average}$ as small as .14 is present in the case where SR_T = .10 and n = 200. This finding clearly illustrates again that the Z_D test had low power when both n and SR_T are small.

Figure 3.5.1 and Figure 3.5.2 present the simulation results where d = 1.00. The probabilities of observing AI for Ln_{adj}, the Z_{IR} test, and the Z_D test were almost the same across most of the conditions in both of these figures. However, statistical power constituted a problem, again, for the Z_D test when both n and SR_T are small. Again, setting SR_T to .10 and n to 200, the Z_D test indicates AI 0% of the time when AR_{min} is .12 and 74% of the time when AR_{min} is .20. Even though there is a large subgroup mean

difference, the results show that the 4/5ths rule fails to indicate AI about half of the time in some cases when $SR_T = .90$. Referring to Figure 3.5.2, for example, the 4/5ths rule indicates no AI about 55% and 50% of the time, respectively for the applicant pool sizes of 200 and 400. Notice that SR_{diff} and $IR_{average}$ are .20 and .79 in these conditions.

The behaviors of Ln_{adj} and the Z_{IR} test are stable across conditions in Figure 3.5.1 and Figure 3.5.2. These two measurements indicate AI almost 100% of the time across all the conditions. The smallest probability for the Z_{IR} test is observed when the $SR_T = .10$ and $n = 200$. In this particular condition the Z_{IR} test indicated AI 86% of the time.

Figure 3.5.1 Simulation Results Where d = 1.00 and AR_{min} = .12

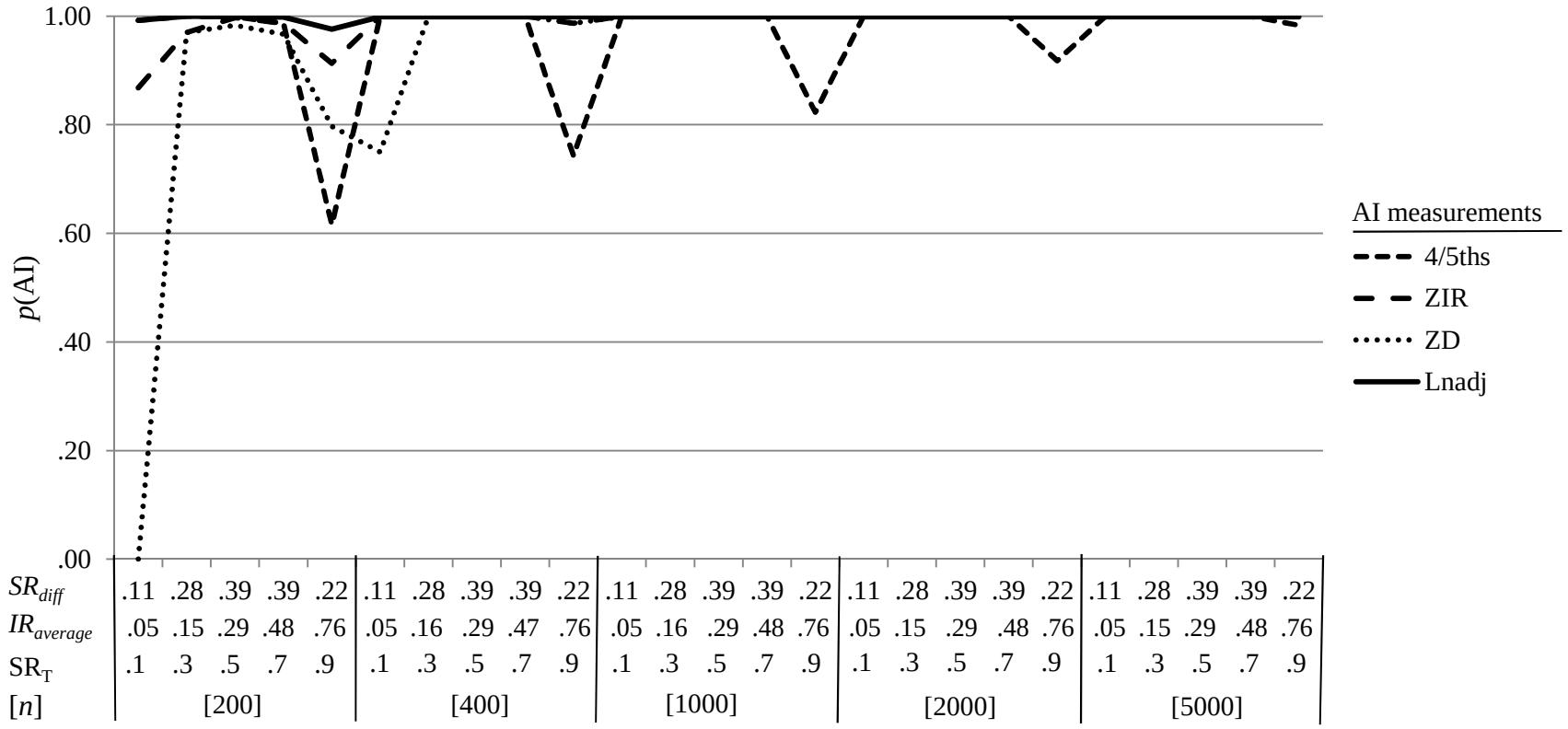


Figure 3.5.1 Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (SR_{diff}), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [n] when d = 1.00 and minority applicant ratio (AR_{min}) = .12.

Figure 3.5.2 Simulation Results Where d = 1.00 and AR_{min} = .20

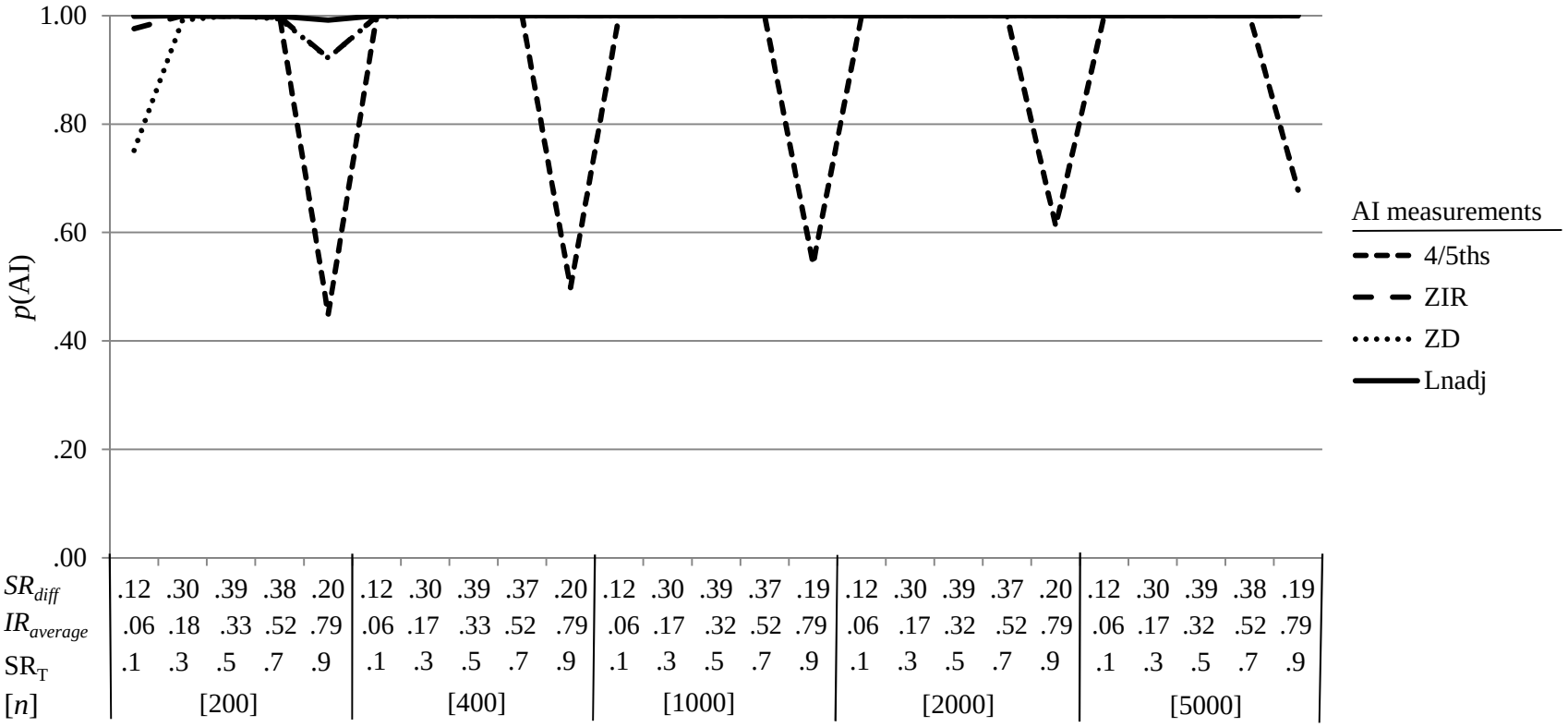


Figure 3.5.2 Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (SR_{diff}), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [n] when d = 1.00 and minority applicant ratio (AR_{min}) = .20.

Appendix B provides the results of simulations for AR_{min} of .30 through Figure 3.1.3 to Figure 3.5.3. The behaviors of these four AI measurements do not deviate much as a response to the increase in AR_{min}. When there is no subgroup mean difference (Figure 3.1.3), a small sample size and SR_T still constitute a problem for the 4/5ths rule and Ln_{adj}, the tests of significance provide acceptable level of Type I error rates across conditions. When there is a small subgroup mean difference (Figure 3.2.3), the behavior of the four AI measures did not change much as AR_{min} increased from .20 to .30. The average probabilities of observing AI remained about the same for the 4/5ths rule and Ln_{adj} and decreased about 6 percent for the Z_{IR} test and 7 percent for the Z_D test. The behavioral patterns of these four AI measurements, however, remain almost the same. The most noticeable change was observed in the behavior of Ln_{adj} in cases where $n \geq 1000$ and $SR_T = .90$. The probabilities of observing AI for Ln_{adj} increased, on the average, about 22 percent on these conditions when AR_{min} was increased from .20 to .30. When there is a moderate subgroup mean difference ($d = .50$), increasing AR_{min} from .20 to .30 did not change the average probabilities of observing AI and the pattern of the four AI measurements very much. A seven percent increase in the average probabilities of observing AI for the Z_D test was the only change worth to mention. When there are large subgroup mean differences ($d = .75$ or 1.00), the behavioral pattern and the average probabilities of observing AI for the all four AI measurements remained almost the same, except that an increase in AR_{min} from .20 to .30 led to an increase in the average probabilities of observing AI for the 4/5ths rule in conditions where $d = 1.00$. All of these observed changes in the behaviors of the four AI measurements as a response to an increase in AR_{min} are minimal and do not affect the material conclusions of this study.

Table 3.1 provides average probabilities of observing AI across 25 conditions (5 SR_T by 5 n) for each of the four AI measurements. These results clearly demonstrate that Ln_{adj} is less likely to indicate AI than the 4/5ths rule in all situations when d is small ($d \leq .25$). Ln_{adj} is also less likely to indicate AI than the tests of significance in three out of the four cases when d is small. The only case where the tests of significance are less likely to indicate AI than Ln_{adj} is observed when $d = 0$ and $AR_{min} = .12$. These results indicate that Type I error rates for Ln_{adj} is, on the average, smaller than those for the 4/5ths rule, the Z_{IR} and Z_D tests.

To make a comparison among the possible Type II error rates for these four AI measurements, average probabilities of observing AI in conditions where $d \geq .50$ were examined. The results reveal that $Ln_{adj,}$ on the average, is more likely to indicate AI than the other AI measurements when there is moderate to large subgroup mean difference. Therefore, Type II error rates for Ln_{adj} are possibly lower than those of the 4/5ths rule, the Z_{IR} test and the Z_D test.

Table 3.1

Average Probabilities of Observing AI for Each AI Measurements

		D				
AR_{min}	AI measurements	0	.25	.50	.75	1.0
.12	The 4/5ths rule	.072	.433	.735	.829	.962
	Ln_{adj}	.069	.430	.918	.986	.998
	The Z_{IR} test	.055	.562	.845	.951	.988
	The Z_D test	.043	.474	.758	.888	.939
.20	The 4/5ths rule	.059	.440	.719	.808	.909
	Ln_{adj} rule	.045	.418	. 934	.993	.999
	The Z_{IR} test	.052	.619	.885	.973	.995
	The Z_D test	.045	.572	.831	.951	.985

Note. Average probabilities were computed by summing all probabilities for each of the pairwise applicant pool size (*n*)-SR_T combinations (5 x 5) and dividing this sum by 25. For example, average probability of observing AI for the 4/5ths rule when AR_{min} is .12 and d value is .25 was computed by summing the observed probabilities graphed in Figure 3.2.1 for this particular AI measurement.

Knowing that OFCCP (2002) recommends using the tests of practical and statistical significance rather than the 4/5ths rule when a large number of hiring is made, it is important to discuss the behavior of Ln_{adj} and compare it to the behaviors of the other discussed AI measurements in cases where the applicant pool size is large. For the reason being that, Table 3.2 provides the average probabilities of observing AI for each of the AI measurements when the applicant pool size is equal to or larger than 1000. As presented, Ln_{adj} keeps average probabilities of observing AI minimum relatively to the 4/5ths rule and the tests of statistical significance when there is small or no subgroup mean

difference. Moreover, the results revealed that Ln_{adj}, on the average, indicates AI at a much greater rate than the 4/5ths rule when there is medium or large subgroup mean difference. This is because, compared to the 4/5ths rule, Ln_{adj} is more sensitive to the changes in subgroup mean difference and less sensitive to changes in SR_T and sample size. These results point out that Ln_{adj} can be considered as an alternative practical significance test to the 4/5ths rule, especially in cases where a large applicant pool size is evident.

Table 3.2

Average Probabilities of Observing AI for Each AI Measurements When n ≥ 1000

		D				
AR_{min}	AI measurements	.00	.25	.5	.75	1.00
.12	The 4/5ths rule	.036	.444	.787	.936	1.000
	Ln_{adj}	.011	.380	.980	1.000	1.000
	The Z_{IR} test	.051	.754	.921	1.000	1.000
	The Z_D test	.043	.695	.967	.998	1.000
.20	The 4/5ths rule	.025	.445	.791	.853	.998
	Ln_{adj}	.003	.356	.991	1.000	1.000
	The Z_{IR} test	.047	.846	.997	1.000	1.000
	The Z_D test	.044	.812	.995	1.000	1.000

Note. Average probabilities were computed by summing all probabilities for each of the pairwise applicant pool size (n)-SR_T combinations (3 x 5) and dividing this sum by 15.

Table 3.3 provides the correlations among SR_{diff}, $IR_{average}$, and the probabilities of observing AI for the four AI measurements. Examining these correlations, one of the

interesting finding is that the correlation between SR_{diff} and Ln_{adj} (r = .766) is larger than both the correlations between SR_{diff} and the Z_{IR} test (r = .729) and between SR_{diff} and the Z_D test (r = .729). This finding is particularly interesting because, knowing that the effect size for the Z_D test is based on the raw selection ratio difference and the effect size for the Z_{IR} test is based on the natural log transformed selection ratio difference, one would easily anticipate otherwise. Not surprisingly, the correlation between SR_{diff} and the 4/5ths rule is relatively small. Nevertheless, considering *SRdiff*, the highest correlation is observed between SR_{diff} and Ln_{adj}. That means, Ln_{adj} is more sensitive to the changes in selection ratio difference than the other AI measurements. Therefore, Ln_{adj} is more likely to indicate AI as the difference between SR_{maj} and SR_{min} increases and less likely to indicate AI as the difference gets smaller. For that reason, if the difference between selection ratios was to be considered as the benchmark to assess AI, Ln_{adj} would outperform the other three AI measurements.

The other benchmark that I used to assess the behaviors of these AI measurements is $IR_{average}$. Exploring the correlations between $IR_{average}$ and AI measurements, the results are in the expected direction. The probabilities of observing AI for each of the AI measurements are negatively related to $IR_{average}$. Among those, the results for the 4/5ths rule has the strongest correlation with $IR_{average}$ (r = -.854). Although not as strong as the correlation observed for the 4/5ths rule, the correlation between the results of Ln_{adj} and $IR_{average}$ (r = -.739) is stronger than the correlations between the results of the tests of significance and $IR_{average}$. The correlations were -.692 and -.601, respectively for the results of the Z_{IR} test and the Z_D test. In short, these correlations suggest that when the

ratio of selection ratios is considered as the benchmark, although not performing as well as the 4/5ths rule, Ln_{adj} outperforms the tests of significance.

In addition to the relatively large correlations with SR_{diff} and $IR_{average}$, Ln_{adj} shows strong correlations with the other AI measurements. Compared to the tests of significance, Ln_{adj} correlated more strongly with the 4/5ths rule (r = .761).Besides, the correlations between Ln_{adj} and the Z_{IR} test (r = .884) and $\mathrm{Ln_{adj}}$ the Z_D test (r = .825) are stronger than the corresponding correlations between the 4/5ths rule and these tests of significance. All these results are in favor of the assumption that Ln_{adj} provides a balance between the 4/5ths rule and the tests of significance.

Table 3.3

Correlation Matrix among Average Selection Ratio Difference (SR_{diff}), Average Sample Impact Ratio ($IR_{average}$), and the Results of the Four AI Measurements

	SR_{diff}	$IR_{average}$	The 4/5ths rule	The Z_{IR} test	The Z_D test	Ln_{adj}
SR_{diff}	1.000					
$IR_{average}$	-.589**	1.000				
The 4/5ths rule	.706**	-.854**	1.000			
The Z_{IR} test	.729**	-.692**	.744**	1.000		
The Z_D test	.729**	-.601**	.674**	.964**	1.000	
Ln_{adj}	.766**	-.739**	.761**	.884**	.825**	1.000

Note. These correlations were computed among the calculated SR_{diff}, $IR_{average}$, and probabilities of AI for the four AI measurements observed in all the simulated conditions (n = 250).

Chapter 4

Discussion

The objectives of this study were to introduce a new practical significance test, as an alternative to the 4/5ths rule, and to compare the behavior of this test to the behaviors of the 4/5ths rule and the two tests of significance: the Z_{IR} test and the Z_D test. As discussed, previous research pointed out some deficiencies with the wholesale application of the 4/5ths rule (Boardman, 1979; Bobko & Roth, 2004; Roth et al., 2006; Shoben, 1978). Furthermore, OFCCP (2002) recommended using tests of practical and statistical significance rather than the 4/5ths rule to assess AI when there is large number of hiring made. That means, not only AI researchers but also employment agencies are aware of the inadequacy of the 4/5ths rule in correctly assessing whether or not a selection procedure causes AI, yet there is no single study proposing an alternative practical significance test. Therefore, proposing an alternative practical significance test, this research is the first attempt to fill in a much needed gap in AI literature.

As the previous research had shown, the main problem observed with the 4/5ths rule is its sensitivity to the changes in SR_T (Roth et al., 2006; Sackett & Ellingston, 1997). On the one hand, this sensitivity causes the 4/5ths rule to produce relatively high Type I error rates when both subgroup mean difference ($d \leq .25$) and SR_T (e.g., .10 and .20) are small. On the other hand, it causes the 4/5ths rule to produce relatively high Type II error rates when SR_T is large (e.g., .80 and .90) and subgroup mean difference is moderate or large ($d \geq .50$). The possibility of high Type II error rates was evident in cases where large

SR_{diff} was observed. Behaving conservatively, the 4/5ths rule failed to provide evidence for AI most of the time in these cases, even though SR_{diff}, as large as .18, was observed.

The deficiency of the 4/5ths rule in identifying AI when SR_T and sample size are large has some potentially devastating impact particularly on the perception of minorities about the fairness of the selection outcome. As the results showed the 4/5ths rule, on the average, indicated no AI 90% of the time when $SR_T = .90$, $d = .75$, and $n \geq 1000$ (see Figure 3.3.1 and 3.3.2). Failing to identify the differential effect of a selection test with a subgroup mean difference of .75 on the selection rates of minority and majority groups, the fitness of the 4/5ths rule as an AI measurement became questionable. A numerical illustration would further help to clarify the discussion. For example, the number of applicants for the Texas Board of Law Exam was 3182 within the last year (July, 2010 and February, 2011) and the passing rate was 84 percent[18]. There were no demographics available about the racial or ethnic background of these applicants. Assuming that racial composition of these applicants was the same as the racial composition of people living in Texas[19], it could be concluded that 45 percent of the applicants were White, 38 percent were Hispanics, 12 percent were Blacks, and 5 percent were from other racial groups. To make the argument clearer, let's change the passing rate to .89. Based on this new passing rate and racial composition given above, a possible selection outcome is presented in Table 4.1.

[18] http://www.ble.state.tx.us/Stats/main_stats.htm
[19] http://quickfacts.census.gov/qfd/states/48000.html

Table 4.1

A Possible Outcome Scenario not Violating the 4/5ths Rule for the Simulated Results of Texas Board of Law Exam

Race	*N* of Applicant	*N* of Hires	*N* of Rejects	Selection Ratio
Whites	1432	1432	0	1.00
Hispanics	1209	968	241	.80
Blacks	382	306	76	.80
Others	159	128	31	.81
Total	3182	2834	348	.89

Using the 4/5ths rule, the results indicate that there is no AI against any of the minority groups, yet 241 Hispanic/Latino applicants out of 1209 and 76 Black applicants out of 382 were failed while none of the 1432 White applicants failed in the test. Regardless of whether or not the 4/5ths rule is violated, it is very hard not only for minority applicants to perceive this outcome as fair but also for the employment agencies and for the organizations to communicate it as a fair outcome. As SR_T approaches 1.00, such scenarios presented in Table 4.1 are to be frequently observed, especially if the selection test has a medium or large *d* value. Behaving conservatively when SR_T is large, the 4/5ths rule mostly fails to indicate AI in these conditions. In this regard, one objective of this study is to propose an AI measurement that behaves less conservatively than the 4/5ths rule in selection scenarios where SR_T is large. As the results indicate, Ln_{adj} meets this objective. Being more likely to indicate AI, Ln_{adj} behaves less conservatively than the 4/5ths rule as the SR_T approaches 1.00. Using the same simulated Texas Board of Law Exam data, Table 4.2 gives the minimum number of applicants to be selected from each

minority groups in order to comply with Ln_{adj}. Here, the number of hires and number of rejects for majority and minority groups are more comparable.

Table 4.2

Minimum Number of Applicants from each Minority Groups Need to Pass Texas Board of Law Exam not to Violate Ln_{adj}

Race	*N* of Applicant	*N* of Hires	*N* of Rejects	Selection Ratio
Whites	1432	1307	125	.91
Hispanics	1209	1055	154	.87
Blacks	382	333	49	.87
Others	159	139	20	.87
Total	3182	2834	348	.89

Another concern with the current AI measurement was their sensitivity to the changes in applicant pool size. This is particularly true for the tests of statistical significance. As the applicant pool size increases, the statistical power of significance tests increases. This increase in the statistical power causes significance test to signal evidence for AI even when there are practically meaningless differences between SR_{maj} and SR_{min}. With an applicant pool size of 5000, for instance, a selection ratio difference of .02 or less will indicate a statistically significant difference. That means, tests of significance behave less conservatively than Ln_{adj} when the applicant pool size is large. Therefore, by behaving more conservatively than the 4/5ths rule and less conservatively than the tests of significance in conditions where both applicant pool size and SR_T are large (see Figure 3.1.1 and 6.2), Ln_{adj}, provides a balance between the current practical and statistical significance tests.

The results also reveal that $Ln_{adj,}$ on the average, produces less Type I error rates than both the 4/5ths rule and the tests of significance, especially when the applicant pool size is equal to or larger than 1000. The Z_{IR} test and the Z_D tests produce low Type I error rates with small *d* values when applicant pool size are 200 and 400. This is the only condition that tests of significance outperform Ln_{adj} by producing low Type I error rates. There is also one particular condition in the simulations where the Type I error rates for the 4/5ths rule are smaller than the Type I error rates for Ln_{adj}. It is the condition where the SR_{min} is set to .12 and applicant pool size is set to 200. Ln_{adj} produces, on the average, lower Type I error rates than the 4/5ths rule in all of the other conditions where the subgroup mean difference set to zero. Therefore, Ln_{adj}, on the average, is more likely to prevent organization from being sued for using a selection test with no AI.

Another important characteristic of Ln_{adj} is that it is less sensitive to the changes in SR_T and applicant pool size and more sensitive to the changes in standardized mean difference between majority and minority groups on the selection test used. That means the standardized group mean difference is the major determinant of the behavior of Ln_{adj}. This helps organizations worry less about variables (SR_T and applicant pool size) over which they do not necessarily have control. The practical effect of this can be clearly understood by focusing on the hypothetical scenarios presented in Table 4.3.

> *An Example Scenario from Table 4.3:* Company A and Company B are going to hire 200 workers for an entry level position. They are planning to use the same off-the-shelf selection test with a *d* value of .25. The mean selection rate for this entry-level position is .70 and minority applicant rate is .12 in the industry. Company A, using various recruitment channels effectively, was able to recruit

more applicants than the industry average. Company B, however, being unsuccessful in the recruitment process, recruited fewer applicants than the industry average. Therefore, the overall selection rate became .60 for Company A and .80 for Company B. After the selection process completed, some of the minorities who were not hired claimed that they were adversely affected by the selection test and filed a formal complaint with EEOC's regional office.

Table 4.3

Changes in the Probabilities of Observing AI for the 4/5ths Rule and Ln_{adj} as the Organizations Become Selective

	N_A	SR_T	*p*(AI) 4/5ths Rule	*p*(AI) Ln_{adj}	*p*(AI) 4/5ths Rule	*p*(AI) Ln_{adj}
			AR_{min} = .12		AR_{min} = .20	
			N_H = 200			
Industry average	286	.70	.28 (.77)	.42 (.90)	.22 (.79)	.44 (.91)
Company A	333	.60	.42 (.89)	.42 (.94)	.42 (.89)	.42 (.94)
Company B	250	.80	.14 (.52)	.41 (.87)	.10 (.55)	.48 (.87)
			N_H = 500			
Industry average	714	.70	.23 (.84)	.47 (.96)	.12 (.85)	.49 (.98)
Company A	833	.60	.41 (.97)	.50 (.98)	.31 (.97)	.44 (.99)
Company B	625	.80	.07 (.57)	.45 (.94)	.02 (.50)	.52 (.98)

Note. N_A = applicant sample size; SR_T = total selection ratio; *p*(AI) 4/5ths rule = probability of observing adverse impact for the 4/5ths rule; *p*(AI) Ln_{adj} = probability of observing adverse impact for Ln_{adj}; AR_{min} = proportion of minority in the applicant pool; and N_H = number of applicants hired. The first probabilities in the table are for the conditions where *d* = .25 and the probabilities in the parentheses are for the conditions where *d*= .50.

Using the 4/5ths rule to measure if there is evidence for AI; EEOC's regional office is 3 times more likely to find evidence of AI for Company A than for Company B in the above scenario. Remember that these two companies used the same selection test and the only difference between them was in their recruitment effort. Company A, being more successful in the recruitment process, became more selective than Company B. As a result of being more selective, Company A increased the expected performance level of those applicants hired (Martin & Raju, 1992) and therefore increased the utility of the selection test[20]. Although being more selective produces a desired outcome in terms of selection utility, it makes organizations vulnerable against AI complaints if the 4/5ths rule is used to measure AI. That means, in a nutshell, the 4/5ths rule discourages organizations from being selective.

EEOC's regional office, however, is almost equally likely to find evidence of AI for Company A and Company B when it uses Ln_{adj}, instead of the 4/5ths rule, to measure AI. The probability of observing AI for company A is .42 and for Company B is .42 in the above scenario when Ln_{adj} is used. The results for other scenarios presented in Table 4.3 demonstrate that Ln_{adj} (compared to the 4/5ths rule) ensures that companies, regardless of being more or less selective, experience similar amount of vulnerability against AI complaints when they employed the same selection test. Thus, Ln_{adj} is fairer than the 4/5ths rule toward organizations. More specifically, Ln_{adj} is almost equally likely to indicate AI for Company B when it indicates AI for Company A and vice versa. Reaching the same conclusion (evidence of AI or no evidence of AI) is particularly important when one thinks about the possible effect of finding evidence of AI for one company and

[20] This is true only if the cost of processing additional applicants is smaller than the gain associated with the increase in the expected performance level of applicants hired as a result of being more selective.

finding no evidence of AI for the other company on the fair competition between them. Returning to the scenario above, Company A can easily lose its customers to Company B as a result of prestige lost due to being sued for using selection test causing AI. Ln_{adj}, however, ensures that the results will be almost the same for the two companies taking the same action (using the same selection test); therefore, it is fairer than the 4/5ths rule.

All things considered, first, Ln_{adj} is less sensitive to the changes in SR_T and applicant pool size; therefore, it provides stable results across conditions. Second, it is more conservative (less likely to indicate AI) than the 4/5ths rule when SR_T is small. That means, unlike the 4/5ths rule, Ln_{adj} will not be violated by a relatively small difference between SR_{min} and SR_{maj}. Third, Ln_{adj} is less conservative (more likely to indicate AI) than the 4/5ths rule when SR_T is larger. In that regard, Ln_{adj} will not fail to indicate evidence for AI when there are selection ratio difference as large as .15 and .18. Fourth, compared to both the 4/5ths rule and the tests of significance, Ln_{adj}, on the average, is less likely to indicate AI when *d* is small and more likely to indicate AI when *d* is moderate or large. All these results indicate that Ln_{adj} is a good alternative not only to the 4/5ths rule but also to the tests of significance, especially when the applicant pool size is large.

Another important contribution of this study was that unlike the previous ones, this study explored the behavior of the tests of statistical significance in conditions where sample size as large as 5000 is evident. Previous research has tended to focus on tests of significance in cases where sample size is small (Collins & Morris, 2008; Morris, 2001; Morris & Lobsenz, 2000), but many of the cases with which OFCCP deals are with large organizations and therefore much larger sample sizes. OFCCP (2002) recommended using tests of statistical significance to assess AI when there is a large number of hiring,

yet rarely has the simulation research evaluated the behaviors of the tests of significance in large samples.

Although it is expected that the null hypothesis (H_0: $\pi_{maj} = \pi_{maj}$) to be rejected when $N \geq 5000$ and $d \neq .00$, exploring the behavior of the tests of significance in cases where large sample size is evident helps researchers to evaluate the recommendation of the OFCCP Manual. As the results presented in this thesis suggest (see Figure 3.2.1 and 3.2.2), the tests of significance indicate AI almost %100 of the time when there is a small *d* value (.25) and large applicant pool size. Indicating AI almost %100 of the time when the effect size is small, the results of the tests of significance should be interpreted with caution. Therefore, I could argue against the recommendation of the OFCCP Manual and state that relying on tests of significance to assess AI when the applicant pool size is equal to or larger than 2000 is not a good alternative to the 4/5ths rule. Furthermore, the OFCCP Manual is not clear what they exactly mean by *very large samples* (italics added). If it means *n* of 1000, complying with OFCCP's recommendation of using tests of significance would be a good consideration. If the manual means *n* of 5000 or larger, however, complying with its recommendation could be argued against. Thus, examining the behavior of the tests of significance in cases where very large sample is evident would help OFCCP and other employment agencies to provide clearer recommendations. The results presented in this thesis will further contribute to the AI research and help employment agencies in this respect.

Furthermore, unlike previous simulations, the current simulation research not only provides the probabilities of observing AI for a selection scenario but also the average selection ratio and average impact ratio for that particular scenario. This helped to

evaluate the behavior of the discussed AI measurements in a more concrete and comprehensive way. Although previous studies discussed that the 4/5ths rule fails to indicate AI when there are large selection ratio difference in cases where SR_T is large, this study explicitly showed how large selection ratio differences were for various selection scenarios in which the 4/5ths rule failed to indicate AI.

In the last place, there have been studies recommending various alternative tests of significance and exploring their behaviors in different selection scenarios (Biddle & Morris, 2011; Collins & Morris, 2008; Morris & Lobsenz, 2000). Although researchers should definitely continue to explore alternative tests of statistical significance, future research should also explore alternative practical significance tests to assess AI. Taking such an approach, as I have in this thesis, would be more in line with the intention of *Uniform Guidelines*.

References

Biddle, D. A., & Morris, S. B. (2011). Using Lancaster's mid-*p* corrections to the Fisher's exact test for adverse impact analyses. *Journal of Applied Psychology.* Advance online publication. doi: 10.1037/a0024223

Boardman, A. (1979). Another analysis of the EEOCC "Four-Fifths" rule. *Management Science, 25*, 770-776.

Bobko, P. & Roth, P. L. (2004). The four-fifths rule for assessing adverse impact: An arithmetic, intuitive, and logical analysis of the rule and implications for future research and practice. *Research in Personnel and Human Resource Management, 23*, 177-198.

Bobko, P. & Roth, P.L. (2010). An analysis of two methods for assessing and indexing adverse impact: A disconnect between the academic literature and some practice. In J.L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 29-49). New York: Routledge.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52*, 561-589.

Cascio, W. F., Outtz, J., Zedeck, S., & Goldstein, I. L. (1995). Statistical implications of six methods of test score use in personnel selection. *Human Performance, 8*, 133-164.

Castaneda v. Partida, 430 US 482 (1977).

Chan, D. & Schmitt, N. (1997). Video-based versus paper-andpencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143-159.

Charge Statistics (n.d.) In U.S. Equal Employment Opportunity Commission Charge Statistics FY 1997 through FY 2011. Retrieved from http://www.eeoc.gov/eeoc/statistics/enforcement/charges.cfm

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Cohen, D. B. & Dunleavy, E. M. (2009). The Center for Corporate Equality releases a review of OFCCP settlements from fiscal year 2007. *The Industrial-Organizational Psychologist, 47,* 145-14.

Collins, M. W. & Morris, S. B. (2008). Testing for adverse impact when sample size is small. *Journal of Applied Psychology, 93,* 463-471.

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining the predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92,* 1380-1393.

Equal Employment Opportunity Commission, Civil Service Commission, U.S. Department of Labor, & U.S. Department of Justice (1978). Uniform guidelines on employee selection procedures. Retrieved from http://uniformguidelines.com/uniformguidelines.html

Equal Employment Opportunity Program (n.d.). Equal employment opportunity terminology. *In National Archives*. Retrieved from http://www.archives.gov/eeo/terminology.html

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26*, 461-477.

Greenberg, I. (1979). An analysis of the EEOCC "Four-Fifths" rule. *Management Science*, *25*, 762-769.

Griggs v. Duke Power Company, 401 U.S. 424 (1971).

Hattrup, K., Rock, J., & Scalia, C. (1997). The effects of varying conceptualization of job performance on adverse impact, minority hiring, and predicted performance. *Journal of Applied Psychology, 82*, 656-664.

Hazelwood School District v. United States, 433 U.S. 299 (1977).

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, *9*, 152-194.

Hsu, L. M. (1993). Using Cohen's tables to determine the maximum power attainable in two-sample tests when one sample is limited in size. *Journal of Applied Psychology*, *78*, 303-305.

Hunter, J. E. (1981). *False premises underlying the 1978 Uniform Guidelines on Employee Selection Procedures: The myth of test validity*. Paper presented to the Personnel Testing Council of Metropolitan Washington, Washington, DC.

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72-98.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. M. (1977). Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring. *Journal of Applied Psychology, 62*, 245-260.

Kirnan, J. P., Farley, J. A., & Geisinger, K. F. (1989). The relationship between recruiting source, applicant quality, and hire performance: An analysis by sex, ethnicity, and age. *Personnel Psychology, 42*, 293-308.

Landy, F. J. (2005). Phases of employment litigation. In F. J. Landy (Ed.), *Employment discrimination litigation: Behavioral, quantitative, and legal perspectives* (pp. 3-19). San Francisco: John Wiley & Sons.

Litigation Statistics (n.d.) In U.S. Equal Employment Opportunity Commission Litigation Statistics FY 1997 through FY 2011. Retrieved from http://www.eeoc.gov/eeoc/statistics/enforcement/litigation.cfm.

Martin, S. L. & Raju, N. S. (1992). Determining cutoff scores that optimize utility: A recognition of recruiting costs. *Journal of Applied Psychology, 77*, 15-23.

Matrixx Initiatives, Inc., et al v. Siracusano et al. 563 U.S. __ (2011).

Morris, S. B. (2001). Sample size required for adverse impact analysis. *Applied HRM Research, 6*, 13–32.

Morris, S. B., & Lobsenz, R. E. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology, 53*, 89-111.

Nuisser, U., Boodoo, B., Bouchard, T. J., Boykin, A. W., Brody, B., Ceci, S. J., … Sternberg, S. J. (1997). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77-101.

Office of Federal Contract Compliance Programs (2002). Office of federal contract compliance manual chapter VII: Identification and remedy of employment discrimination. Retrieved from http://www.dol.gov/ofccp/regs/compliance/fccm/chptr7.pdf

Ployhart, R. E. & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*, 153-172.

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9*, 241-258.

Pyburn, K. M., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity-validity dilemma: Overview and legal context. *Personnel Psychology, 61*, 143-151.

Ricci et al. v. DeStefano et al. 557 US.__ (2009).

Roth, P. L., Bobko, P., & Switzer III, F. S. (2006). Modeling the behavior of the 4/5ths rule for determining adverse impact: Reasons for caution. *Journal of Applied Psychology*, *91*, 507–522.

Rumelt, R. (1991). 'How much does industry matter?' *Strategic Management Journal, 12*, 167-185.

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology*, *50*, 707-721.

Sackett, P. R., & Roth, L. (1996). Multi-stage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. *Personnel Psychology, 49,* 549-572.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56,* 302-318.

Schmidt, F. L., Greenthal, A. L., Hunter, J. E., Berner, J. G., & Seaton, F. W. (1977). Job sample vs. paper-and-pencil trades and technical tests: Adverse impact and examinee attitudes. *Personnel Psychology, 30,* 187-197.

Shoben, E. W. (1978). Differential pass-fail rates in employment testing: Statistical proof under Title VII. *Harvard Law Review, 91,* 793-813.

Theron, C. (2009). The diversity-validity dilemma: In search of minimum adverse impact and maximum utility. *SA Journal of Industrial Psychology, 35,* 183-195.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 9,* 165-181.

Upton, G. J. G. (1982). A comparison of alternative tests for the 2 × 2 comparative trial. *Journal of the Royal Statistical Society, 145,* 86-105.

U.S. Census Bureau (2010). USA quick facts. Retrieved from http://quickfacts.census.gov/qfd/states/00000.html

U.S. Department of Labor & U.S. Bureau of Labor Statistics (2011). Labor force characteristics by race and ethnicity, 2010. Retrieved from http://www.bls.gov/cps/cpsrace2010.pdf

Appendix A: The Simulation Code Used

```
rep =1000 # number of iterations (constant)
sr =.1   # selection ratio (variable)
pmin =.2     # minority ratio (variable)
nap =1000   # total number of applicant (variable)
nsle = sr*nap            # number of applicant selected (variable)
nb = nap*pmin            # number of minority applicants (variable)
nw = nap-nb            # number of majority applicants (variable)
mb = 100            # minority mean (variable)
sb = 15             # minority standard deviation (variable)
mw = 100              # majority mean  (constant)
sw = 15               # majority standard deviation (constant)
vctr <- c(1:rep)        # create a vector for the result of simulations
for(i in 1:rep){        # open a loop function
db = rnorm(nb,mb,sb)           #minority applicant sample
dw = rnorm(nw,mw,sb)          # majority applicant sample
db.t <- t(rbind(db,rep(1,nb)))  #minority applicant sample and indicator 1
dw.t <- t(rbind(dw,rep(0,nw))) #majority applicant sample and indicator 0
d <- rbind(db.t,dw.t)          # combining minority and majority applicant sample
N1 <- c(d[,1])              # nesting the data in a data frame and assigning column labels
N2 <- c(d[,2])
X <- c(1,2)
df <- paste("N",seq(along=x),sep="")
A <- data.frame(lapply(df,get))
names(A) <- df
sort1.A <- A[order(N1),]   # ordering the data according to column 1
y <- sort1.A[(nap-nsle+1):nap,] # trimming applicants who are not selected
V1 <- c(y[,1]) # nesting the trimmed data in a data frame and assigning column labels
V2 <- c(y[,2])
z <- c(1,2)
hr <- paste("V",seq(along=z),sep="")
B <- data.frame(lapply(hr,get))
vctr[i] <-sum(V2) # sum the values in column-2 and nest the sum in the vector created
}               # close the loop function
min.h <- vctr            # number of minority hired
maj.h <- nsle-min.h       # number of majority hired
p.min <- min.h/nb         # ratio of minority hired
p.maj <- maj.h/nw         # ratio of majority hired
ff.rule <- p.min/p.maj        # 4/5th rule
```

Appendix A: The Simulation Code Used (continued)

```
ff <- ff.rule<.80        # simulations resulted with AI according to 4/5th rule
r1 <- sum(ff)            # number of simulations with AI for the 4/5
p.ff <- r1/rep           # probability of obtaining AI when 4/5th rule is used
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
z1 <- log(p.min/p.maj)
z2 <- 1-sr
z3 <- (nap)*(sr)*(pmin)*(1-pmin)
z4 <- z2/z3
z5 <- sqrt(z4)
z6 <- z1/z5
z7 <- abs(z6)
zir <- z7>1.96           # simulations resulted with AI for the Z_IR test
r2 <- sum(zir)           # number of simulations with AI for the Z_IR test
p.zir <- r2/rep          # probability of obtaining AI for the Z_IR test
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
u1 <- p.min-p.maj
u2 <- sr*(1-sr)
u3 <- nap*pmin*(1-pmin)
u4 <- u2/u3
u5 <- sqrt(u4)
u6 <- u1/u5
u7 <- abs(u6)
zd <- u7>1.96            # simulations resulted with AI for the Z_D test
r3 <- sum(zd)            # number of simulations with AI for the Z_D test
p.zd <- r3/rep           # probability of obtaining AI for the Z_D test
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
q1<-log(p.maj)/log(p.min)
q2<-log(2.63+.8*sr)
q3<-q1*q2
ln.adj<-q3<.8            # simulations resulted with AI for Ln_adj
r4<-sum(ln.adj)          # number of simulations with AI for Ln_adj
p.ln.adj<-r4/rep         # probability of obtaining AI for Ln_adj
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
y1<-sum(ff.rule)         # sum of the impact ratios
ir.av<-y1/rep            # average impact ratio
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
y2<-sum(u1)              # sum of the selection ratio difference
```

Appendix A: The Simulation Code Used (continued)

```
sr.diff<-y2/rep          # average selection ratio difference
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
sr.diff
ir.av
p.ff
p.zir
p.zd
p.ln.adj
```

Appendix B: Simulation Results When AR_{min} = .30

Figure 3.1.3 Simulation Results Where *d* = .00 (no AI) and AR_{min} = .30

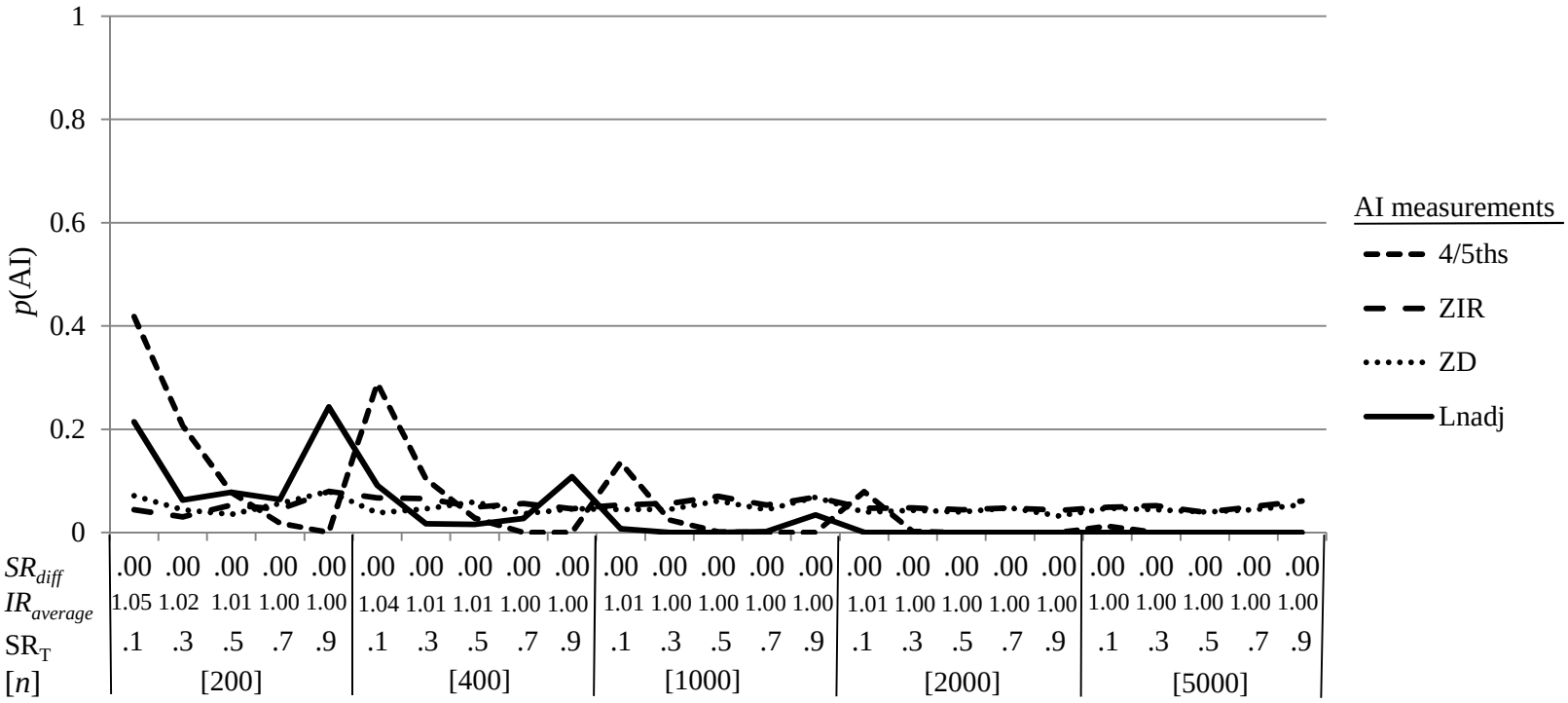


Figure 3.1.3 Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (SR_{diff}), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [*n*] when *d* = .00 and minority applicant ratio (AR_{min}) = .30.

Appendix B: Simulation Results When AR_{min} = .30 (continued)

Figure 3.2.3 Simulation Results Where d = .25 and AR_{min} = .30

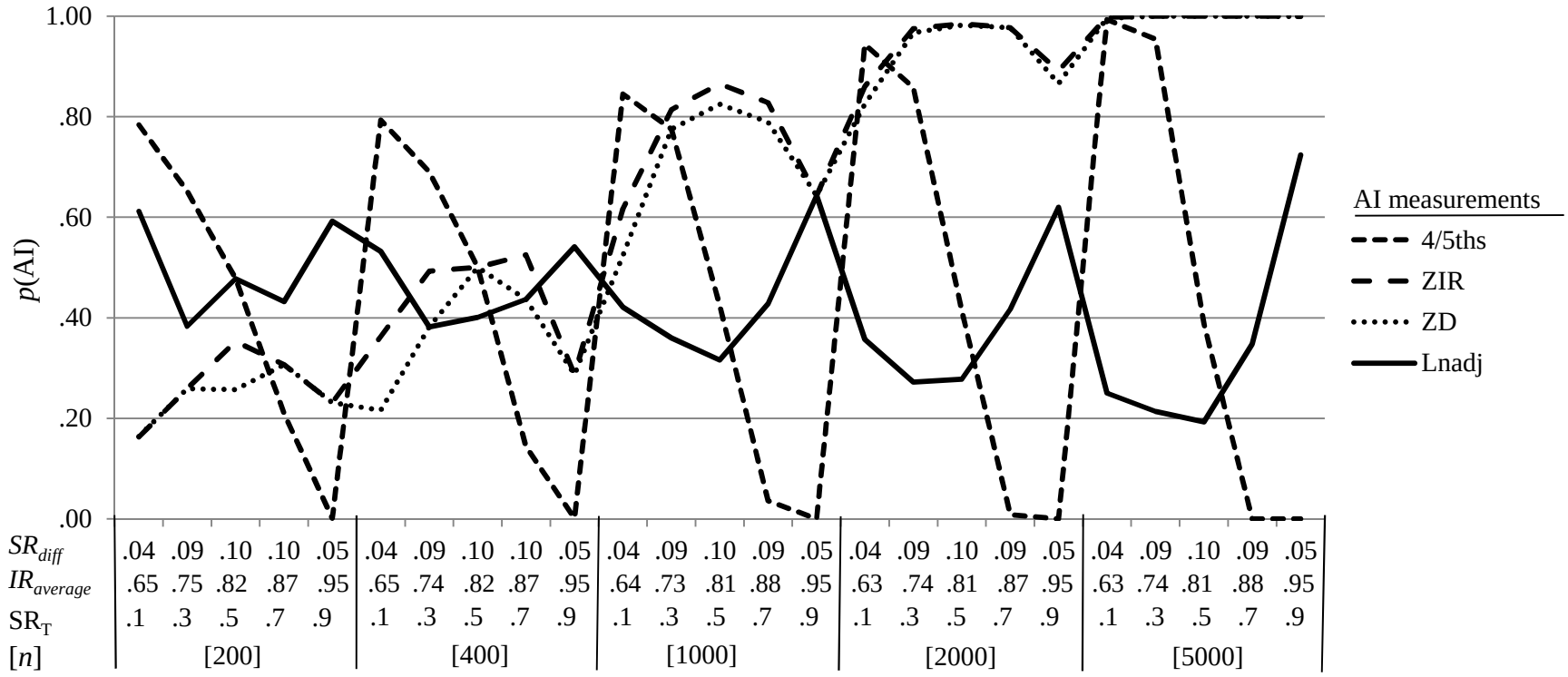


Figure 3.2.3 Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (SR_{diff}), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [n] when d = .25 and minority applicant ratio (AR_{min}) = .30.

Appendix B: Simulation Results When AR_{min} = .30 (continued)

Figure 3.3.3 Simulation Results Where *d* = .50 and AR_{min} = .30

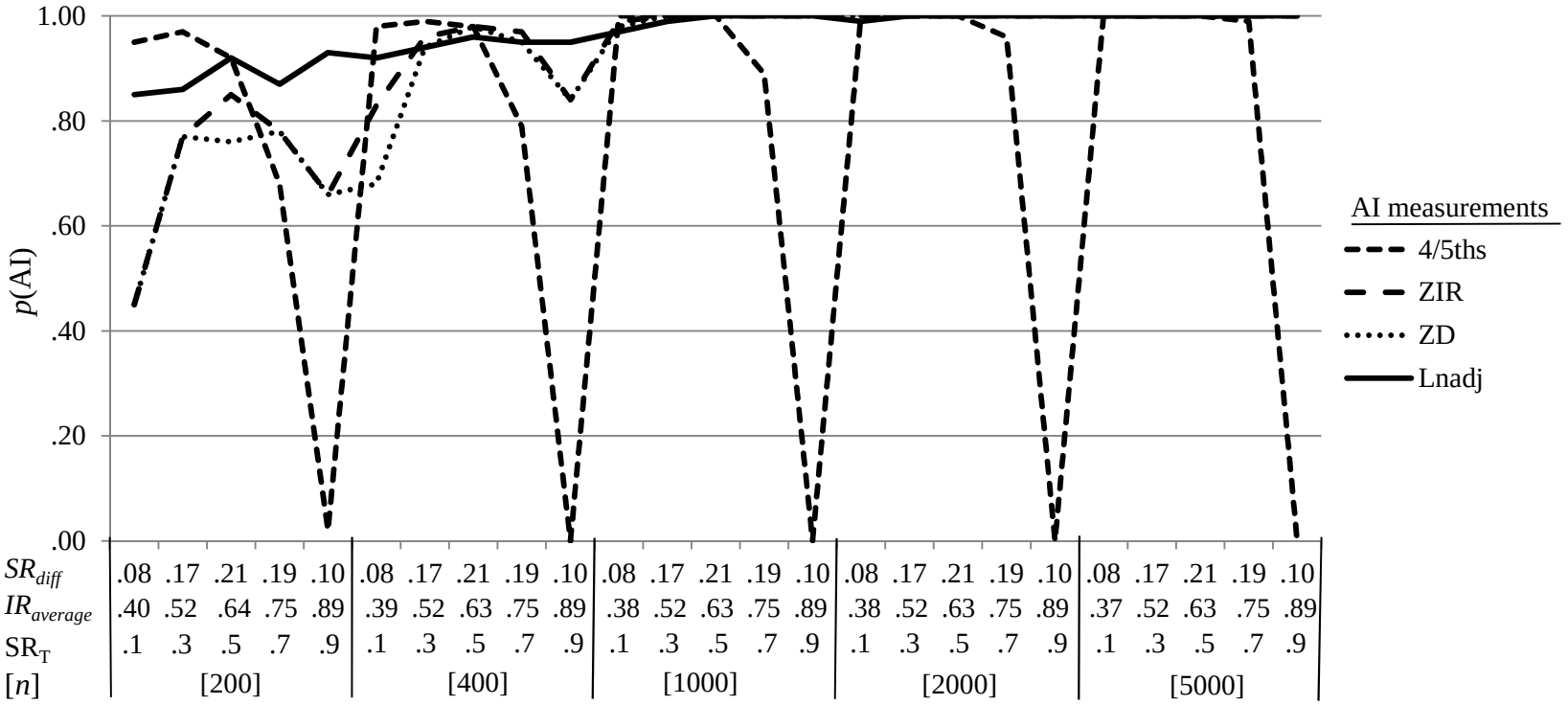


Figure 3.3.3 Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (SR_{diff}), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [*n*] when *d* = .50 and minority applicant ratio (AR_{min}) = .30.

Appendix B: Simulation Results When AR_{min} = .30 (continued)

Figure 3.4.3 Simulation Results Where *d* = .75 and AR_{min} = .30

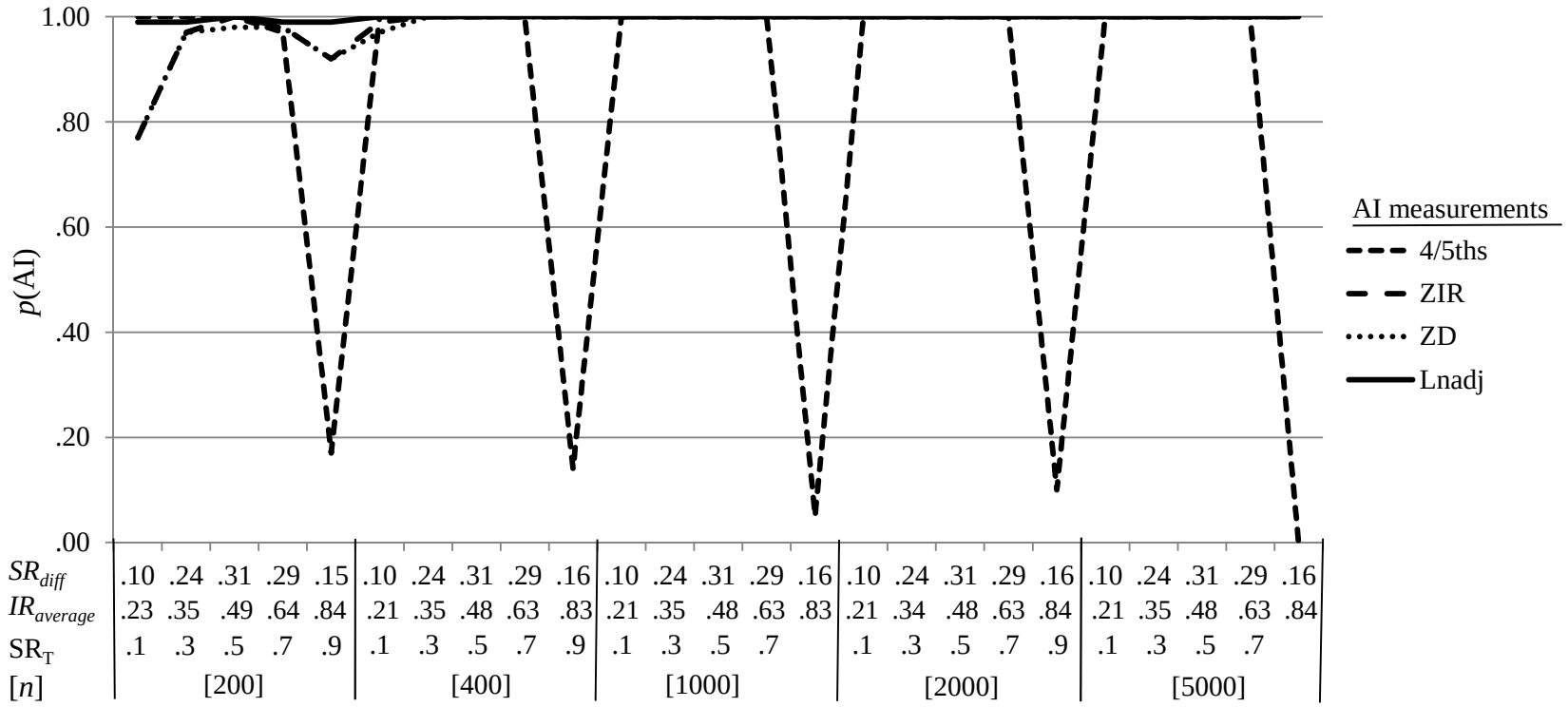


Figure 3.4.3 Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (SR_{diff}), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [*n*] when *d* = .75 and minority applicant ratio (AR_{min}) = .30.

Appendix B: Simulation Results When AR_{min} = .30 (continued)

Figure 3.5.3 Simulation Results Where *d* = 1.00 and AR_{min} = .30

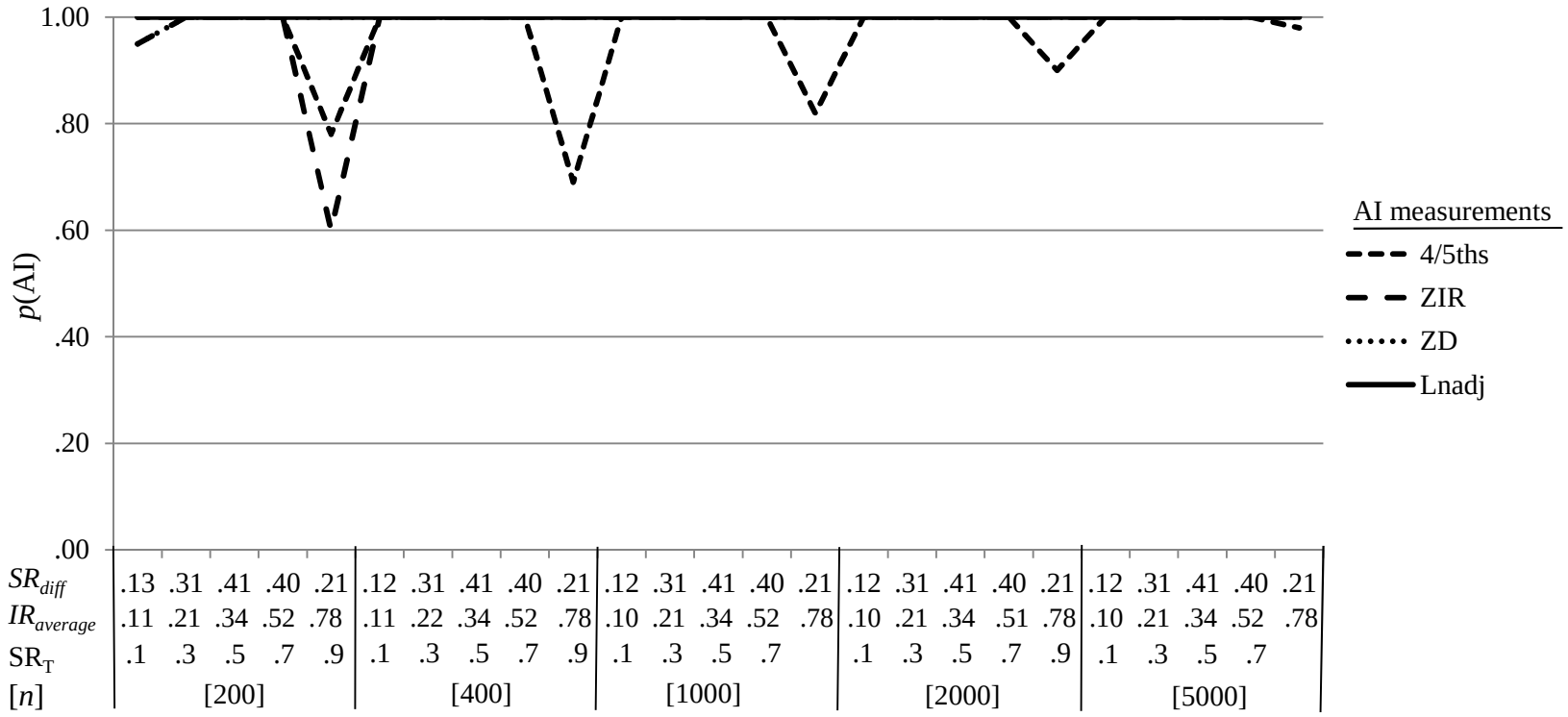


Figure 3.5.3 Variations in the probabilities of observing AI for each AI measurements by average difference between SR_{maj} and SR_{min} (SR_{diff}), average sample impact ratio ($IR_{average}$), total selection ratio (SR_T), and applicant pool size [*n*] when *d* = 1.00 and minority applicant ratio (AR_{min}) = .30.